

UNIVERSITÄT
DUISBURG
ESSEN

Open-Minded

DOCTORAL THESIS

**Harnessing Latent Space Semantics for Enhanced
Interpretability of Recommender Systems in Item
Retrieval and Online Communication Dynamics**

by Tim Donkers

Harnessing Latent Space Semantics for Enhanced Interpretability of Recommender Systems in Item Retrieval and Online Communication Dynamics

Von der Fakultät für Informatik
der Universität Duisburg-Essen

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte kumulative Dissertation

von

Tim Donkers

aus

Bad Neuenahr

1. Gutachter: Prof. Dr.-Ing. Jürgen Ziegler

2. Gutachter: Univ.-Prof. Dipl.-Ing. Dr. Dietmar Jannach

Tag der mündlichen Prüfung: 12. Januar 2024

Abstract

Recommender systems, in today's digital age, have emerged as influential algorithmic curators, significantly shaping content consumption, e-commerce, and public opinion. While the majority of research in this area has been anchored in algorithm development and offline evaluation, user-centered considerations are still conspicuously neglected. In particular, pervasive technologies, while powerful, tend to operate as inscrutable *black boxes*, concealing their inner workings from both developers and end users. By exploring the potential for improving the interpretability of the latent information spaces employed by many recommendation models, this work emphasizes the importance of understanding the structural conditions that form the underlying data base. It highlights the depth of insight that can be gained from the intricate relationships between the entities under consideration, and aims to bridge the current gaps in our understanding of recommender systems.

The first goal of this thesis is to advance model-based recommender systems by harnessing the power of latent information spaces. To this end, two novel methods are introduced: Tag-enhanced Matrix Factorization (TagMF) and Aspect-based Transparent Memories (ATM). TagMF extends traditional matrix factorization by intertwining associations between items and tags, thereby indirectly inferring user-tag relationships as well. This not only enriches the user-item preference matrix, improving prediction accuracy and mitigating data sparsity issues, but also increases the degree to which the otherwise latent semantics can be interpreted. On the other hand, ATM leverages user reviews and applies deep learning techniques to provide evidence-backed recommendations. By associating the semantic subtleties within its latent space with concrete user utterances, ATM paves the way for transparent recommender systems that more closely resemble how humans justify their evaluations of items. Together, these approaches, validated by user studies, enable users to interactively navigate and influence the recommendation process, increasing both perceived self-efficacy and recommendation quality.

The second objective introduces a methodology grounded in simulation, offering a nuanced lens to comprehend social phenomena pertinent to recommender systems in online social networks. Recognizing the gaps in existing research, this approach underscores the significance of psychological and sociological factors in deciphering the impact of these systems. Traditional offline evaluations, predominantly centered on predicting item ratings or rankings, often bypass the diverse influences on decision-making within recommendation tasks. While user-centric experimental studies have sought alignment between recommendation technology and psychological or demographic attributes, they tend to narrow down phenomena to individual experiences. In environments like social networks, where collective dynamics are crucial, broader effects must be considered to fully grasp the societal implications of recommender systems. To bridge these research voids, our methodology harnesses the rich semantic connections inherent in latent information spaces. It seeks to analyze phenomena such as the polarization of ideological groups by tracing the evolution of these semantics. This perspective allows us to understand how the dynamics between individual users, their social environment, and algorithm-driven content distribution together influence the spread of opinions and the manifestation of particular beliefs within online networks.

In summary, this thesis presents an alternative approach to studying recommender systems that emphasizes human factors and latent semantics. It argues that, beyond predicting ratings or

ranking alternatives, additional information needs can be satisfied by accessing and disclosing latent semantics inferred by contemporary machine learning technologies, ultimately enabling users to properly evaluate system-side suggestions. In addition, the work proposes a simulation paradigm to reconcile psychological and sociological factors influencing decision-making and communication behavior with variations in recommendation technology. These contributions aim to improve our understanding of the complex relationship between users and recommender systems, ultimately enhancing their functionality and impact.

Keywords Recommender Systems, Collaborative Filtering, Machine Learning, Deep Learning, Latent Space Semantics, Interpretability, Transparency, User-centered Design, Online Social Networks, Social Polarization, Filter Bubbles, Echo Chambers, Systems Theory, Simulation, Opinion Dynamics, Agent-based Modeling.

Acknowledgements

The journey of this dissertation has been both challenging and rewarding, made possible by the encouragement, guidance, and camaraderie of many. It is a reflection not only of my efforts, but of the collective spirit of inquiry and support of my mentors, colleagues, and loved ones.

I am deeply grateful to my advisor, Jürgen Ziegler, whose belief in my abilities and generous academic freedom have significantly shaped my research trajectory. His encouragement allowed me to pursue my interests and fostered a nurturing atmosphere for my intellectual growth. His insightful feedback has been a valuable guide, allowing me to refine my ideas and push the boundaries of my understanding.

I would also like to express my sincere appreciation to Benedikt Loepp, my mentor since my undergraduate studies and later a cherished colleague. His reliability, supportive nature, and cooperative attitude have not only enriched my academic journey, but also exemplify the essence of a dedicated researcher. His quality of engaging me at eye level, fostering a space for collaborative inquiry, and his steadfast assurance have been inspiring and vital to my confidence as a scientist.

I am eternally thankful to my family, whose unwavering support has been my stronghold. My mother, Marita Donkers, has been a beacon of unconditional love and patience at every stage of my life. Her tolerance for my quirks, evidenced by the countless times she retrieved my forgotten gym bag from the school bus, has provided a comforting assurance that has allowed me to fearlessly explore my own path. My sister, Anna Donkers, has also been a source of comfort and belief in me, providing a foundation of love and encouragement that has propelled me forward on this journey.

Last but certainly not least, my friends have been an incredible source of encouragement, offering warmth, insightful discussion, and unyielding support during difficult times. Their belief in me and the deep conversations we have shared have been instrumental in shaping my ideals and interests.

Finally, this acknowledgment is a small token of my gratitude to everyone who has contributed to my academic journey. Their influence extends beyond the confines of this dissertation and leaves a lasting imprint on my academic and personal growth. Thank you.

— *Tim Donkers, October 2023*

Contents

1	Introduction	1
2	Content-enhanced Collaborative Filtering	5
2.1	General Overview of Collaborative Filtering	5
2.2	Tag-enhanced Matrix Factorization	8
2.2.1	Content Integration	9
2.2.2	Interaction Framework	10
2.2.3	Latent Semantics	14
2.2.4	User Study	14
2.2.5	Limitations & Future Directions	15
2.3	Aspect-based Transparent Memories	17
2.3.1	Recommendation & Argumentation	17
2.3.2	Transparent Memories	19
2.3.3	Aspect-based Interpretability	20
2.3.4	Terminological Delineation	22
2.3.5	User Study	24
2.3.6	Limitations & Future Directions	25
2.4	Conclusion	25
3	Online Social Networks & Ideological Social Systems	27
3.1	The Dynamics of Recommender Systems & User Segregation	27
3.2	Group Identity & Differentiation	30
3.2.1	In-group & Out-group	31
3.2.2	Friend-Foe Dichotomy	32
3.2.3	Delineation of Echo Chambers	34
3.2.4	Derivation of Ideological Social Systems	35
3.3	Ideological Social Systems	36
3.3.1	Systems-theoretical Classification	37
3.3.2	Structural Coupling	42
3.4	Conceptual Framework	46
3.4.1	Psychic & Social Systems	47
3.4.2	Recommender Systems	53
3.5	Simulation Framework & Empirical Validation	57
3.5.1	Literature Review & Model Comparison	58
3.5.2	Simulation Procedure & Validation	60
3.6	Conclusion	69
4	Contributions	73
	Bibliography	79

List of figures

96

List of tables

97

Papers Included in the Cumulus

Content-Enhanced Collaborative Filtering

Donkers, T., Loepf, B., & Ziegler, J. (2017). Sequential User-Based Recurrent Neural Network Recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (pp. 152-160). Association for Computing Machinery. <https://doi.org/10.1145/3109859.3109877>

Donkers, T., Loepf, B., & Ziegler, J. (2018). Explaining Recommendations by Means of User Reviews. In A. Said, & T. Komatsu (Eds.), *Joint Proceedings of the ACM IUI 2018 Workshops* (Vol. 2068). RWTH Aachen. <http://ceur-ws.org/Vol-2068/exss8.pdf>

Loepf, B., Donkers, T., Kleemann, T., & Ziegler, J. (2019). Interactive recommending with Tag-Enhanced Matrix Factorization (TagMF). *International Journal of Human-Computer Studies*, 121, 21-41. <https://doi.org/10.1016/j.ijhcs.2018.05.002>

Donkers, T., Kleemann, T., & Ziegler, J. (2020). Explaining recommendations by means of aspect-based transparent memories. In F. Paternò, & N. Oliver (Eds.), *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 166-176). The Association for Computing Machinery. <https://doi.org/10.1145/3377325.3377520>

Donkers, T., & Ziegler, J. (2020). Leveraging Arguments in User Reviews for Generating and Explaining Recommendations. *Datenbank-Spektrum*, 20(2), 181-187. <https://doi.org/10.1007/s13222-020-00350-y>

Online Social Networks & Ideological Social Systems

Donkers, T., & Ziegler, J. (2021). The Dual Echo Chamber: Modeling Social Media Polarization for Interventional Recommending. In *Proceedings of the 15th ACM Conference on Recommender Systems* (pp. 12-22). Association for Computing Machinery. <https://doi.org/10.1145/3460231.3474261>

Donkers, T., & Ziegler, J. (2023). De-sounding echo chambers: Simulation-based analysis of polarization dynamics in social networks. *Online Social Networks and Media*, 37-38, 100275. <https://doi.org/10.1016/j.osnem.2023.100275>

Introduction

Recommender systems have become pervasive tools in the digital landscape, serving as algorithmic curation mechanisms that sift through vast amounts of information to personalize and organize content in various online contexts. These systems have a significant impact on online content consumption, e-commerce, or the formation of public opinion. Platforms ranging from streaming services to online social networks rely on recommendation technology to deliver personalized experiences, ensuring content relevance and increasing user engagement.

While recommender systems are deeply embedded in our digital experiences, our understanding of how they influence content choices, purchasing behaviors, social interactions, and the information we access remains incomplete. This knowledge gap is exacerbated by the fact that the delicate workings of these systems are often obscured, making them inaccessible to both developers and end users. As recommender systems increasingly rely on machine learning and deep learning techniques, the resultant models, while powerful, tend to operate as inscrutable *black boxes* [DLZ17; Zha*19; RA20; Son*21]. The opacity of these systems makes it challenging to understand and interpret the underlying selection processes, leading to difficulties in identifying biases or other undesirable consequences, such as perpetuating stereotypes [Ado13; Che*23] or reinforcing filter bubbles [Par11; Ngu14; Jia*19]. Users may struggle to comprehend the rationale behind recommendations, resulting in reduced trust in the system and decreased satisfaction with the suggested content [TM07; Zha18; Pec19]. The lack of contextual information or explanations for recommendations can cause users to feel overwhelmed or disoriented, especially when confronted with a plethora of options.

Additionally, technologically determined filter bubbles can inadvertently create echo chamber effects, primarily exposing users to content that aligns with their existing preferences, consequently reducing the diversity encountered [DV*15; Tö18; Bri*20; TBB21; DZ21; DZ23]. This can contribute to opinion polarization and hinder users from discovering novel or serendipitous content. Hence, the algorithms behind these systems, often operating in the background, play a pivotal role in determining what information is presented to users, thereby influencing their decisions, preferences, and even beliefs. As these systems become increasingly integrated into our daily digital interactions, their role in influencing individual choices and collective behaviors cannot be understated.

Yet, despite these challenges and shortcomings, the analysis of recommendation technologies in both academia and industry has focused primarily on technological aspects, with most research

efforts taking place in restrictive offline environments [BL15; Bee*16; Faz18]. New approaches are typically developed and assessed in isolation, using historical data to drive iterative production cycles. While empirical datasets provided by established platforms are often employed for internal validation [e.g. HM16; HK15; BM*11], they primarily serve to confirm a model’s performance against predefined metrics and baseline models. This process, while efficient, tends to prioritize quantitative measures such as prediction accuracy, serving as a mere proxy for the actual needs people might have when using these systems. By deliberately transforming behavioral signals into data points and making predictions based on them, outcomes are primarily examined within the confines of tight boundaries, inadvertently widening the gap between the results of offline evaluations and the applicability of a model when deployed in diverse, real-world scenarios.

In light of the challenges posed by the opacity of these systems, there has been a growing emphasis on making recommendations more transparent and interpretable. Rather than prioritizing predictive performance, explainable recommender systems aim to give users insight into how recommendations are generated, fostering deeper understanding and trust [ZC*20; CZW22]. This can be achieved through various techniques, such as conveying feature importance [Loe*19], employing case-based reasoning [McS05; DKZ20], generating aggregated visualizations of the underlying data [KLZ17], or even offering the possibility to converse with an automated agent [HBZ23]. By making the recommendation process more transparent, users can gain a better understanding of why a particular item was suggested, which can potentially increase their satisfaction with the system and enhance the overall user experience.

Alongside the increasing attention to transparency, there has been a significant shift towards user-centered evaluations [PCH11; Kni*12; KR12]. Beyond the scope of explainable systems, these evaluations are essential for understanding the wide range of user interactions with recommender systems. Specifically, they address ethical dimensions like trust [Har*14; Kun*19], bias [Che*23], and fairness [Eks*22; Del*23], as well as user experience factors such as diversity [KP17; CHV22], serendipity [GDBJ10; KWV16], and user autonomy [MTF20; Var20]. Moreover, they consider the broader psychological [Bol*10; Kni*12], economic [LH19], and social impacts [DZ21; RD22; TM22; DZ23] that these systems have on users. Consequently, advanced empirical methods – such as user studies or A/B testing – which focus on explicitly involving humans in the loop to adopt a more holistic perspective, have become increasingly vital components of evaluation projects. Notably, experimental laboratory studies immerse participants in controlled environments, linking recommendation technology to psychological or demographic variables. By placing a stronger emphasis on human experience, these studies provide a more rounded understanding of the interaction between users and technology than data-driven metrics alone.

While some research endeavors have ventured into the territory of user studies, they often tend to narrow down their focus to individual experiences. Such micro-level analyses, though valuable, can sometimes miss the broader social dynamics at play, especially in platforms like online social networks. The collective behaviors, influenced by recommender systems, can lead to larger societal patterns and ramifications, such as echo chambers or ideological polarization. Moreover, determining long-term effects of human-machine interaction through laboratory settings is difficult, as most studies are limited to single or a few consecutive sessions. Longer-term studies tend to be costly and produce knowledge at a significantly slower pace. Additionally, the number of study participants is often limited, and while crowdsourcing websites such as Amazon

Mechanical Turk¹ or Prolific² facilitate participant recruitment, they introduce their own set of limitations [Sim13; KKH15; Lov*18].

Given the drawbacks of the available repertoire of analytical and empirical tools, as well as the limitations imposed by the black-box nature of many technologies, we recognize the need to broaden the perspective in the analysis of recommender systems. Latent information spaces, which form the basis of many advanced model-based recommendation methods [KBV09; Zha*19], serve as a crucial bridge between raw data and meaningful insights in this regard. These spaces, derived from vast amounts of data, capture intricate relationships and patterns that may elude overt analysis. Yet, although these models may appear opaque, they are central to this thesis due to the immense expressiveness they internally retain [AC09; Mik*13; Loe*19]. Their importance lies in their ability to represent complex user-item interactions in a condensed form that captures the semantic relationships between the entities under consideration. By tapping into the geometry of these latent spaces, we not only address the challenges posed by their initial opacity, but also leverage their rich expressiveness to advance our research goals.

Bearing this in mind, this thesis is structured into two main parts, each addressing a distinct aspect of recommender systems. Our first aim is to harness latent semantics to refine model-based recommender systems. By integrating the otherwise unintelligible dimensions extracted from behavioral data with additional content features, we elevate the transparency and interpretability of the recommendations generated. We propose that connecting learned latent contexts with human-readable information, such as tags or user-written reviews, will help us elucidate the decision logic the system applies to suggest content. This increased level of interpretability, importantly, also helps to implement system features that allow users to interactively control the recommendation process, ultimately improving both perceived self-efficacy and recommendation quality.

Secondly, beyond the realm of recommendation methodology in the narrow sense, this thesis ventures into the domain of simulations to analyze the broader psychological and sociological phenomena associated with recommender systems, especially in contexts like online social networks. By harnessing the semantic richness of latent information spaces, our simulation methodology seeks to represent complex social interactions and emergent patterns in a novel manner, diverging from traditional analytical [e.g. HK02; Def*00] or simulation models [e.g. Sas*21; DGL13] of opinion dynamics. This approach not only provides insights into critical societal phenomena, such as ideological group polarization, but also bridges the research voids between individual user experiences, collective social dynamics, and the influence of recommendation technology on these processes. This enables us to develop novel tools of analysis that extend beyond mere accuracy assessments, as well as address research questions that are challenging to explore through user-centered approaches like experimental laboratory studies.

As powerful algorithmic curators, recommender systems have undeniably reshaped the digital landscape, influencing content consumption and shaping public opinion. However, as this thesis underscores, there is an urgent need to move beyond mere algorithmic efficiency and delve deeper into the human-centric aspects and broader societal implications. By highlighting the interpretability of latent information spaces, this thesis illustrates how increasing the transparency

¹<https://www.mturk.com/>

²<https://www.prolific.co/>

of these models not only demystifies their inner workings, but also provides a clearer understanding of their pervasive influence in our daily lives. Exploring the dynamics of human-technology interaction through simulation helps us elucidate how recommender systems shape individual choices while influencing collective behavior and communication dynamics. Through these contributions, this thesis aims to improve our understanding of the complex relationship between users and recommender systems, ultimately improving their functionality and societal impact.

Content-enhanced Collaborative Filtering

Collaborative filtering, a foundational method in many recommender systems, primarily relies on historical user preferences to predict future interests [Sch*07; SK09; KRB21]. However, its reliance on traditional feedback mechanisms, often limited to ratings, highlights a pressing challenge: the need for improved transparency, interactivity and more comprehensive feedback mechanisms. This limitation often results in a lack of nuanced understanding of user preferences. Furthermore, the inherent lack of transparency in many collaborative filtering methods raises concerns, especially when users seek to understand the rationale behind recommendations. While other obstacles, such as sparse user feedback and the dynamic nature of user preferences, persist, the need for enhanced interactivity and transparency stands at the forefront of the investigations undertaken in this thesis.

In light of these challenges, this chapter introduces two novel approaches to collaborative filtering. The first method, Tag-enhanced Matrix Factorization (TagMF) [Loe*19], integrates user-generated tags into the matrix factorization [KBV09; TB22] process, aiming to provide a nuanced understanding of user preferences. The second, Aspect-based Transparent Memories (ATM) [DKZ20], leverages user reviews and deep learning techniques to enhance recommendation quality while offering insights into the recommendation rationale.

The primary objectives of this chapter are to address the inherent limitations of collaborative filtering and to present methodologies that enhance the accuracy, transparency as well as the interactivity of recommender systems. Through a detailed exploration of these methods, we aim to contribute to the ongoing discourse on personalized recommender systems and their role in the digital landscape.

2.1 General Overview of Collaborative Filtering

Automated recommender systems, exemplified by platforms like Amazon [SL17], Twitter [Twi23], YouTube [CAS16], or Netflix [GUH15; Ste*21] primarily operate on the principle of personalization. They use their operationalization of individual preferences, needs, and goals to tailor recommendations specific to each user. Notably, the non-personalized approach, where all users receive identical recommendations, for instance based on general popularity, has been largely overshadowed due to its limited effectiveness in the current digital landscape. While the latter may still yield acceptable results in certain contexts, it is personalized recommender systems

that have come to the forefront, both in industry and academia, due to their superior efficacy. However, it is crucial to acknowledge that personalized systems bring with them their own set of challenges. They are inherently more complex, demanding individual representation for each user and a sophisticated understanding of their unique behavior.

The cornerstone of personalized recommendation generation commonly lies in harnessing the collective *wisdom of the crowd*. One popular approach known as collaborative filtering takes advantage of the aggregated feedback from the entire user community, operating on the assumption that the historical preferences of users with tastes similar to a target user can serve as an effective predictor of that user's future interests [Sch*07]. Collaborative filtering essentially creates a bridge between past user-item interactions and future user behavior, leveraging the patterns discerned within these interactions to generate personalized recommendations.

However, a significant challenge faced by collaborative filtering-based systems is the inherent sparsity of user feedback. To illustrate, consider the dataset released for the Netflix Prize, a prestigious competition held in the late 2000s aimed at enhancing Netflix's recommendation algorithm [BK07]. This dataset was characterized by an overwhelming 98.8% sparsity, highlighting the extent of the problem. This scarcity of feedback means that, for the vast majority of user-item pairs, no explicit feedback is available. This transforms the problem of collaborative filtering into a complex exercise of filling in the blanks. It essentially becomes a task of predicting the missing feedback, i.e., estimating how a user would rate or interact with items they have not yet encountered. Overcoming this issue requires sophisticated techniques and methodologies to make accurate predictions in the face of such widespread data scarcity.

A key advantage of relying solely on user feedback for the relevant items is that developers do not need to ensure the availability of predefined meta-data or extensive expert knowledge. Instead, predictive models can be built using only readily accessible data provided by the end users themselves. This approach does present issues, such as the cold-start problem [LKH14; GJ17; Wei17], but research and real-world applications have demonstrated that collaborative filtering can deliver highly accurate predictions in many cases. In fact, despite the underlying technology being several years old, collaborative filtering remains the industry standard for recommender systems.

Collaborative filtering methods are fundamentally divided into two categories: memory-based (or neighborhood-based) techniques [Sch*07] and model-based approaches [KBV09]. Memory-based methods, embodying the principle of the *wisdom of the crowd* quite literally, make predictions by identifying similarities between users or items. This principle is realized as these methods generate automatic assumptions about the interests of a user by collecting preferences or taste information from many users. The underlying assumption is that if a person A has the same or a similar opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person. This collective decision-making process often leads to better outcomes than could be made by any single user. The simplicity and straightforwardness of these methods facilitate their application across diverse domains and enable rapid algorithmic adaptation. However, their performance is contingent upon the availability of data. In cold-start scenarios, such as the introduction of a new item or the registration of a new user, the initial absence of relational information impedes the determination of user- or item-side neighborhoods. Additionally, the necessity to store all user-item interactions for neighborhood computation poses significant memory and computational challenges.

In response to these limitations, model-based techniques have gained prominence, demonstrating substantial improvements in algorithmic performance and scalability [KRB21]. These techniques initially require high-dimensional user-item interactions, but subsequently employ dimension reduction methods, such as matrix factorization [TB22], to identify a smaller set of dimensions that capture the essence of these interactions. This approach is predicated on the assumption that hidden patterns exist in user feedback behavior, enabling the calculation of user similarities based on their preferences. Similarly, items can be compared as the extracted dimensions represent shared properties among them. Although the dimension reduction process is computationally intensive and memory-demanding, it can be performed in advance, yielding significant performance gains compared to neighborhood-based methods.

Deep learning has emerged as a transformative force in various domains, from image processing to natural language processing, due to its ability to automatically discover salient structures in large datasets. In the realm of recommender systems, deep learning techniques, especially recurrent neural networks³, offer the potential to model temporal dynamics of user behavior, capturing sequential patterns and long-term dependencies in consumption sequences. This is particularly advantageous as it allows the system to predict not just based on static user preferences but also on evolving consumption patterns. Notably, our contributions to this field have been centered around harnessing the power of recurrent neural networks [DLZ17]. Incorporating distinct characteristics of the recommender systems domain, specifically through explicit user representations, our methodology seeks to harmoniously merge individual consumption idiosyncrasies of a user with the overarching temporal patterns observed in item consumption sequences across the broader user base. This not only enhances the capabilities of model-based recommender systems but also ensures they are adaptive to the dynamic nature of user preferences.

However, the computational efficiency and potential improvement in prediction quality, especially those offered by deep learning-enhanced model-based techniques, come at the cost of system transparency and intelligibility. The architectures of deep networks, while powerful, are particularly complex, making it challenging to decipher the rationale behind their recommendations. While the approach detailed in [DLZ17] offers some level of insight by allowing inspection of attention values [BCB14], the inherent opaqueness of the latent representations remains a challenge. In contrast, simpler methods, such as neighborhood-based techniques, offer outputs that are more geometric and readily interpretable, highlighting the trade-off between performance and transparency in modern recommender systems.

Despite these challenges, latent space models remain prevalent in both academic and industrial settings, driven by the advent of big data and rapid advancements in computational technology. These developments have opened new avenues for enhancing recommender systems, particularly through the incorporation of additional content features such as tags [TSMST08; AAJ22] or user reviews [CCW15; GM21]. This approach, termed here as *content-enhanced collaborative*

³It is worth noting that while recurrent neural networks have significantly impacted recommender systems research, other deep learning architectures and techniques have also made substantial contributions. For instance, convolutional neural networks [LeC*98] have been employed for image-based recommendations [TPH19], and transformers [Vas*17] have been utilized for sequence-based recommendation tasks [Xia*21]. Autoencoders [ZLJ20] and graph neural networks [Wu*22] have further enriched the landscape of deep learning in recommender systems. However, a comprehensive overview of the literature in this domain is beyond the scope of this work. For an extensive review, readers are referred to [Zha*19]

filtering, augments the existing framework with content-based information, thereby generating more nuanced and contextually relevant recommendations.

In this chapter, we explore two novel approaches in this vein: Tag-enhanced Matrix Factorization (TagMF) and Aspect-based Transparent Memories (ATM). TagMF extends traditional matrix factorization by integrating associations between items and tags and inferring a relationship between users and tags. This enriches the user-item preference matrix, providing a more detailed understanding of users' preferences, improving prediction accuracy, and addressing the sparsity challenge.

Meanwhile, ATM utilizes user reviews as supplementary data, employing deep learning techniques, specifically neural memories [WCB15; Suk*15]. This approach enhances the quality of recommendations and improves transparency by providing supportive evidence or explanations for recommended items. ATM illuminates the semantic relationships within the latent space of the model, moving towards more interpretable and accountable recommender systems.

By delving into these two approaches, this chapter demonstrates the evolution of collaborative filtering in response to traditional limitations. We highlight the potential of content-enhanced recommender systems in delivering more accurate, transparent, and user-friendly recommendations, marking a promising exploration into the future of personalized recommender systems.

2.2 Tag-enhanced Matrix Factorization

In order to enhance the interpretability of model-based collaborative filtering systems, we propose an approach that seeks to uncover latent patterns within the available feedback data. This approach is predicated on the assumption that machine learning techniques can be harnessed to discern generalizable statistical relationships between the entities under study. In line with the widely accepted notion that latent dimensions derived from user interactions correlate with real-world concepts, we posit that these dimensions can be made interpretable by enriching them with additional information.

Pioneering research on the quantitative representation of semantic contexts has demonstrated that properly trained latent spaces are organized in a way that positions semantically similar entities in close proximity [Zha*06; AC09; Mik*13]. Consequently, the geometric relationships among entities permit an interpretation of their semantic connections⁴. This level of transparency can be valuable to both developers and end users in gaining a better comprehension of a system's behavior. Specifically, it allows for an estimation and delineation of the different concepts captured by each model dimension, significantly simplifying the understanding of how a model makes sense of its underlying data landscape.

Crucially, this interpretative access to the model's functioning not only enhances transparency but also opens avenues for improving the overall user experience. By granting users access to this interpretable layer, they can directly influence the model's functioning without needing to fully comprehend the system's intricacies. This approach offers a nuanced and user-centric way of interaction, thereby addressing key limitations of many cutting-edge recommendation methods. While these are highly adept at identifying suitable items with minimal interaction effort

⁴See Section 3.3.2 for a detailed investigation of the connection between latent space geometry and semantics.

and cognitive load, this high level of automation often results in users feeling constrained in their ability to flexibly express their current interests or access more diverse or novel recommendations [FH09b; Kni*12]. The limited number of items recommended inevitably makes it challenging to maintain an overview of the entire item space. Furthermore, there are significant limitations on how individuals can influence the recommendation process. In many instances, users have no means to do so, even though perceived control plays a crucial role in evaluating a recommendation situation and significantly contributes to user satisfaction.

Specifically, in typical commercial systems utilizing collaborative filtering, the only avenue for users to actively influence the result set is by providing explicit feedback through ratings for individual items. While this allows for some influence, it hinders the situational expression of search interest. As such, it may also intensify the filter bubble effect, as only the long-term profile is refined and alternatives are gradually excluded from the potential recommendation space [Par11; Jia*19]. Furthermore, it is worth noting that explicit feedback is not the only parameter considered by modern recommender systems. Research has demonstrated that implicit feedback, i.e., the surplus of behavior automatically produced during system usage, can often generate more accurate recommendations than explicit rating behavior [HKV08]. However, from the user's perspective, this further complicates developing a solid understanding of the system and maintaining control over adapting recommendations to situational needs.

2.2.1 Content Integration

From this perspective, we have identified a need for models that offer enhanced interpretability and interactive control without compromising predictive power. The performance of the recommender systems, which are based on a model that factorizes the user-item matrix, has been extensively validated and confirmed. Consequently, these systems serve as the foundation for our investigation.

Within the realm of model-based methods, numerous attempts have been made to integrate additional contextual information [e.g. WB11; DG*15; ZNY17]. However, these efforts primarily aim to enhance the predictive performance, often overlooking the user-centric perspective. The predominant focus on model quality results in minimal effort being allocated to intertwining the additional data with latent factors in a manner that maintains the learned correlations post-training. Consequently, the contextual content merely serves as an auxiliary input that may enhance accuracy, but is processed in a manner that obscures its impact on specific recommendations.

In contrast, a multitude of interactive recommendation methods exist that mainly rely on input data beyond the standard user-item feedback [HPV16]. These methods liberate users from the necessity of providing ratings, thereby offering them enhanced control over the recommendation process.

One such method is the interactive adjustment of preference weights, exemplified by the Taste-Weights system [BOH12]. This hybrid recommender system empowers users to manipulate the individual influence of preferred artists, data sources, and associated entities on music recommendations through interactive slider-weights. The system visually highlights the connections between the user profile, data sources, and results, thereby enhancing the transparency of the recommendation generation process. The added interaction capabilities of this system have

been shown to significantly enhance perceived recommendation quality and overall user satisfaction.

Another method of interest is the critique-based recommendation, which enables users to iteratively refine the items shown as recommendations by applying critiques [CP12]. This approach is grounded in the assumption that users find it easier to critique existing recommendations than to articulate a precise search goal from the outset, especially when their domain knowledge is limited. Users can critique items based on one or more characteristics they wish to emphasize or diminish, leading to recommendations that are similar but better align with these refined preferences.

Interactive methods, such as interactive weight adjustment or critique-based recommendation, provide a rich source of inspiration for enhancing the interactive control capabilities of model-based recommender systems. Our goal is not merely to improve interpretability but to extract additional value from the data by integrating concepts from these methods. Specifically, we aim to leverage the sophistication of collaborative methods to address both the user's initial elicitation of preferences within a system during cold start scenarios and to enable interactive control options in subsequent stages.

In order to merge benefits of both model-based and interactive recommendation methods, we build upon an approach proposed by [FZ11], which enhances a latent information model based on matrix factorization with content attributes. This method maintains the relationships between these attributes and the latent factors post-training due to the regression-constrained formulation of the optimization problem. This allows the content-based association to be utilized in the user interface for practical purposes, such as visually representing the learned semantics or providing interactive user access to the model via critiquing and weighting.

In summary, our goal in enriching model-based collaborative filtering with content information goes beyond just improving recommendation accuracy. We aim to use the available data to create interactive mechanisms to influence recommendations, providing a level of control that transcends what standard item-level feedback allows. Our main objective is to improve user control and experience by combining item-related information with collaborative feedback and latent representations, allowing users to effectively adjust their position in the latent information space.

2.2.2 Interaction Framework

Collaborative filtering systems face several challenges when dealing with user feedback on items. As already mentioned, the limitations of rating-based mechanisms often make it difficult for users to express their diverse information needs through such constrained interaction options. Furthermore, this feedback is most effective and relevant when it reflects users' long-term preferences. Considering that user profiles typically gather data over several weeks, months, or even years, it becomes essential to offer mechanisms that allow users to intervene in the filtering process to articulate short-term objectives.

This is where our strategy of integrating additional content becomes pertinent. The deep integration of item-related contextual information with latent factors can be harnessed to implement weighting or critiquing mechanisms. Specifically, users can assign weights to different concepts represented in the model, which are then applied to the values of the corresponding dimensions

in the latent representation. This allows users to indicate a temporary preference for results that incorporate more or less of one or even a combination of multiple concepts. As such, this multifaceted intervention mechanism offers a significantly higher level of expressiveness compared to the abstract evaluation of items based on ratings. However, it is important to note that, as our method fundamentally relies on collaborative technology, it still allows users to express their preferences through ratings and ensures that the result set remains personalized.

The expressiveness of these interactive features largely relies on the selection of appropriate content information. We have, therefore, centered on the exploitation of user-generated tags, which offer several advantages compared to potentially abstract expert knowledge or other meta-data: Firstly, user-generated data is often more easily accessible. Secondly, tags embody concepts in the language used by users themselves, making them inherently understandable and describing objects at a 'local' rather than 'global' level. In other words, they are unique to each user rather than universal like other attributes. Thirdly, tags have previously demonstrated significant potential for enhancing transparency and, crucially for us, controllability [GJG14]. However, only a few approaches have integrated tags directly into latent factor models, and when they have, performance was only assessed in offline experiments [TSMST08; AAJ22].

For these reasons, we employ tags as a running example to investigate the impact of content enhancement, not only for implementing new interaction mechanisms but also more broadly regarding user experience. Notably, we do not need prior knowledge about the relevance of these tags for the current user: our extensions to Forbes and Zhu's original method enable us to estimate preference tendencies for each user without them ever having added tags themselves. As our method paves the way for a wide array of interactive mechanisms, rather than permitting only one type, users can independently select their preferred interaction form, particularly those typically unavailable within collaborative filtering environments. Finally, it is worth mentioning that although we use tags for the current case, the approach is flexible enough to potentially accommodate other attributes due to its minimal data requirements.

In the following, we will briefly explore potential applications of our framework, demonstrating how our method, in conjunction with user-generated tags, can enhance user interaction at various stages of the recommendation process. This includes preference elicitation during a cold start, recommendation adjustments in response to situational needs, and critiques of specific items. Most importantly, we will illustrate how our approach can increase the transparency of the information space and help identify the semantics embedded in individual user profiles.

Preference Elicitation The first application addresses the challenge of initial preference estimation in collaborative systems. Traditional systems require the user to rate a certain number of items before accurate interest prediction can occur. Our content-based approach, however, can leverage feedback from other users as input data. Thus, a new user needs only to provide a few tags, similar to content or knowledge-based methods, to create a user representation within the underlying factor space without requiring ratings. While ratings can be added at any time, users are not obliged to supply tags independently; information can be obtained through alternative means such as social media profiles. Consequently, users can immediately receive recommendations that are internally connected to the dimensions represented by the tags.

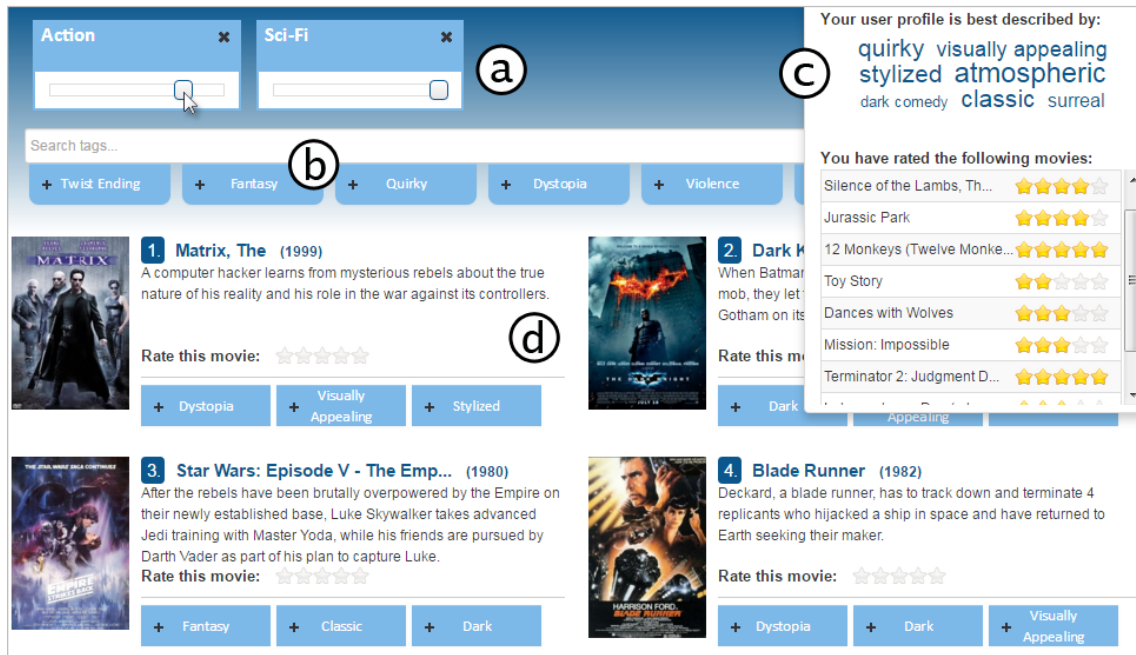


Figure 2.1 Screenshot of the prototype RS for the first user study: The current user has weighted the tags “Action” and “Sci-Fi” (a), therefore receiving matching movie recommendations such as “Matrix” or “Star Wars” (d). The user can also search for other tags provided by the users or get inspiration from the suggestions (b). Furthermore, the user’s existing profile is explained by a tag cloud (c).

Recommendation Adjustment The second application focuses on controlling the ongoing recommendation process. Traditional collaborative filtering systems may fail to meet evolving or situational user needs as they primarily create long-term profiles. Our content-based approach allows users to modify their profile according to the content-based associations with the latent factors or to seek alternative suggestions for more diverse or novel results (Figure 2.1). A user-tag vector exists within the latent space that can be manipulated through interactive assignment of weights to tags. This enables real-time, continuous adaptation of the initial recommendation set based on the user’s long-term profile, allowing users to interactively explore the effects of their preference settings and escape potential “filter bubbles”. In principle, we are no longer predicting actual ratings. Instead, we combine the user’s general preference structure with the operationalization of their current interests or situational (e.g., mood- or activity-based) needs, which are expressed with respect to the tags by interacting with the system.

Critiquing Our approach also enables users to interact with model-based CF systems in a more discrete manner, akin to the critiquing approach [CP12]. Similar to the MovieTuner system, users can request items that are overall similar to a currently recommended item, but differ in representation of selected dimensions (see Figure 2.2). This allows for the consideration of specific context-dependent or situational aspects of the search and decision process. For example, if the movie *Apocalypse Now* is shown, a user might critique it as “too dark”, leading to a recommendation of *Saving Private Ryan*.

Movies based on **Apocalypse Now** and your user profile.

This movie **I want...**

war less equal more
 dark less equal more
 horror less equal more
 romance less equal more
 aliens less equal more
 sci-fi less equal more
 comedy less equal more

Search tag... + Add

Recommendation

Galaxy Quest
Decades after the success of the sci-fi series "Galaxy Quest," the show's washed-up stars -- Jason Nesmith, Gwen DeMarco and Alexander Dane -- are unwittingly recruited by actual aliens to pull off an intergalactic rescue mission.

Relevant tags: + space + satire + funny
Choose as movie to critique

Back to the Future
Eighties teenager Marty McFly is sent back in time to 1955, inadvertently disrupting his parents' first meeting and attracting his mother's romantic interest. Marty must repair the damage to history by rekindling his parents' romance...

Relevant tags: + time travel + adventure + classic
Choose as movie to critique

Life Is Beautiful
A touching story of an Italian book seller of Jewish ancestry who lives in his own little fairy tale. His creative and happy life would come to an abrupt halt when his entire family is deported to a concentration camp during World War II. While ...

Relevant tags: + bittersweet + world war ii + drama
Choose as movie to critique

The Intouchables
A true story of two men who should never have met - a quadriplegic aristocrat who was injured in a paragliding accident and a young man from the projects.

Relevant tags: + drama + based on a true story + true story
Choose as movie to critique

Dylan Moran: Like, Totally...
Dylan Moran, creator of Channel 4's BAFTA award-winning "Black Books" returns with an all-new stand-up show. Unpredictable, startling, bizarre, elegiac, but above all brilliant and hilariously funny, Moran is a master of comedy.

Relevant tags: + funny + satire + drama
Choose as movie to critique

Your user profile is best described by:
 fantasy disturbing thought-provoking atmospheric space war drama funny time travel alternate reality aliens horror true story thriller sci-fi adventure

You have rated the following movies:
 Back to the Future Part III ★★★★★
 Independence Day ★★★★★
 Ace Ventura: Pet Detective ★★★☆☆
 The Blair Witch Project ★★★★★
 The Mummy ★★★★★
 Charlie's Angels ★★★★★
 Star Wars: Episode II - Attack of the Clones ★★★★★
 Grease ★★★★★
 Titanic ★★★★★
 Dumb and Dumber ★★★★★

Figure 2.2

Screenshot of the prototype RS for the second user study: A user whose profile is shown in the dialog (c) has applied a critique (a) to the currently recommended movie “Apocalypse Now” using the tags “dark” and “comedy”. As a consequence, recommendations that fit to the critique and to his or her long-term interests are shown (d). To add further critique dimensions, the user can also search for tags provided by other users (b).

Our method, grounded in matrix factorization, also leverages the user’s long-term profile. This means that the recommendations generated after a critique are not only related to the critiqued item (generally similar, but differing according to the critique), but also consider the user’s overall interests, as reflected in their past user-item interactions. This is a standard practice in collaborative filtering recommenders. For instance, a user who generally prefers comedies might be recommended the movie *M*A*S*H* instead of *Saving Private Ryan* when searching for war movies. Furthermore, the latent information available through TagMF can subtly influence the critiquing process. This allows the resulting recommendations to reflect nuanced item characteristics that explicit tag data alone might not capture.

2.2.3 Latent Semantics

Finally, beyond improving interactive user experience, it is feasible to gain insights into the importance of each dimension within a latent factor space and its connection to the content attributes being examined. By employing dimensionality reduction techniques, such as eigendecomposition [GVL13], we can analyze the attributes that load most positively or negatively on each factor to acquire a basic understanding of the semantics derived from the automatically trained matrix factorization. For instance, Table 2.1 displays the content associations that arise when training 20 factors on a complete dataset and associating them with the 20 most popular tags. The rows represent factors in descending order of importance, with the tags listed in the remaining columns. The matrix's positive and negative values are displayed in each cell, signifying the direction and strength of these relationships, thereby indicating how strongly specific characteristics expressed by the tags are represented in each factor. This enables the embedded semantics of the factors to be individually determined.

This greatly simplifies interpreting the semantics of the latent dimensions. For example, factors 4 and 5 appear to be closely associated with the *fantasy* tag, with factor 4 also demonstrating a negative and factor 5 a positive relationship with the *action* tag. Consequently, both factors correspond to different types of fantasy films. In the rightmost column, we also provide sample films for each factor by identifying the highest-ranked items when all other factors are excluded. Specifically, these are films with at least 10,000 in box office revenue and the highest loadings for the corresponding factor in the item-factor matrix. Accordingly, *The Wizard of Oz* (factor 4) and *Star Wars: Episode IV - A New Hope* (factor 5) clearly align with the previous observations.

At a more general level than these factor representatives, latent space semantics also enable us to illuminate the geometric relationships between users and items. Specifically, they allow us to determine the encaptured meaning of the specific latent dimensions. Regarding items, Figure 2.3 displays an example with two tags that we utilized to train a simple two-factor model. The left plot depicts movies as a function of their (normalized) tag relevance scores, which were extracted from the so-called tag genome. In other words, these scores are based on the row vectors of the item-attribute matrix. Conversely, the right plot arranges the films according to their (also normalized) latent factor values, i.e., based on the vectors of the item-factor matrix. By comparing these two plots, we can observe that the similarities between items in terms of their content attributes remain discernible within the latent information space.

2.2.4 User Study

The external validation of the TagMF approach was conducted through a user study involving 46 participants. It was designed to assess the impact of TagMF on user experience, recommendation quality, and the ability of users to interact with the system. The study was conducted using a web-based prototype movie recommender system, with two variants: one with a standard matrix factorization algorithm, and the other implemented based on TagMF. The interface of the TagMF variant was extended with several tag-based interaction mechanisms, allowing users to manipulate their preferences in real-time. Participants were asked to interact with the system, rate items, and adjust their preferences using the tag-based interaction mechanisms. The system then generated recommendations based on the user's long-term profile and their current interests, as expressed through their interactions with the tags.

Table 2.1 Example for automatically learned relationships between latent factors (rows) and user-generated tags (columns): The five most important factors are shown together with positively (yellow) and negatively (purple) related tags. The factor importance is depicted in brackets in the left-most column. Representatives for each factor are automatically determined by extracting the movies (with at least 10 000 ratings) that score highest for the respective factor in the actual item-factor matrix.

Factor	action	atmospheric	based on a book	classic	comedy	dark comedy	disturbing	dystopia	fantasy	funny	psychology	quirky	romance	sci-fi	surreal	time travel	thought-provoking	twist ending	violence	visually appealing	Automatically extracted representatives
1 (1.66)	0.25	0.38	-0.14	0.47	-0.20	0.16	0.14	0.04	-0.27	-0.15	-0.09	0.17	-0.26	-0.03	0.15	-0.19	0.06	-0.36	0.11	0.24	The Shining, Taxi Driver, A Clockwork Orange
2 (1.51)	-0.11	-0.12	0.02	-0.34	0.12	0.21	0.26	0.12	0.27	-0.13	-0.02	0.14	-0.30	0.22	0.36	-0.51	-0.06	0.09	0.14	-0.21	Natural Born Killers, Brazil, Beetlejuice
3 (1.30)	0.10	0.11	-0.13	-0.63	-0.10	0.11	-0.06	0.07	-0.16	0.06	-0.07	0.21	-0.03	-0.16	0.18	0.17	-0.11	-0.05	-0.07	0.59	Amélie, Sin City, Magnolia
4 (1.21)	-0.39	-0.06	0.24	0.12	-0.16	0.00	0.03	-0.12	0.50	-0.22	0.05	0.14	0.29	-0.17	0.17	0.02	-0.08	-0.04	-0.48	0.19	Wizard of Oz, Willy Wonka & the Chocolate Factory, The NeverEnding Story
5 (1.17)	0.44	0.17	0.01	0.13	-0.11	-0.29	-0.16	0.01	0.44	0.10	-0.12	-0.27	-0.42	0.15	0.02	-0.16	0.01	-0.05	-0.20	0.28	Star Wars: Episode IV – A New Hope, Hobbit: An Unexpected Journey, Thor: The Dark World

The results of the study indicated that TagMF was successful in improving the user experience and the quality of recommendations. Participants expressed positive feedback, and the scores for constructs such as mean item rating and choice satisfaction were better for TagMF in all cases. Furthermore, the integration of tags led to improvements in transparency, resulting in users being more satisfied with their choices. Interestingly, the study also highlighted that high recommendation accuracy does not necessarily translate into equivalent user satisfaction. While conventional approaches at times tend to over-specialize, recommending too many similar items, the use of TagMF resulted in significant or at least marginal improvements regarding the diversity of recommended item sets.

The study also found that the use of latent information in the TagMF approach had a positive impact on the critiquing process and the resulting recommendations. The recommendations not only related to the critiqued item but also took into account the user's general interests based on past user-item interaction data, a combination which apparently led to a satisfying interactive experience. However, the study did not find a positive effect of applying TagMF concerning choice difficulty. Although it was assumed that considering the user's preference profile would make it easier to decide, the results were only slightly better for TagMF.

2.2.5 Limitations & Future Directions

TagMF, while innovative and promising, is not without its potential limitations and challenges. One of the primary constraints is its dependence on item-related tag relevance information. The quality and completeness of this data can significantly impact the effectiveness of the approach. Inaccurate or incomplete tag relevance information could potentially compromise the performance of the method, leading to less accurate or less relevant recommendations. Furthermore,

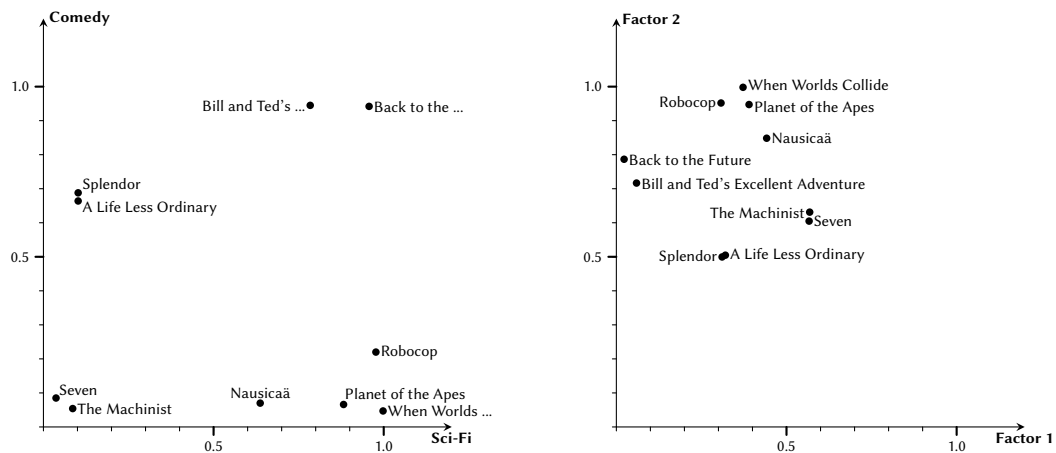


Figure 2.3 Normalized movie positions with respect to tag relevance (left) and latent factors (right).

the complexity of establishing user-tag relations presents another challenge. TagMF attempts to create a relationship between users and tags, which can be a complex task as users' interests in certain tags may not be static; they can change over time and vary depending on the context. This dynamic nature of user-tag relations can make it difficult to accurately capture and represent these relations, potentially affecting the accuracy of the recommendations generated.

Interestingly, our user study found that the application of TagMF did not necessarily simplify the decision-making process for users. While the method may improve the relevance of recommendations, it does not inherently make it easier for users to choose from the set of recommendations. This suggests that the method, while enhancing the quality of recommendations, may not necessarily streamline the decision-making process for users.

In terms of transparency, the improvements achieved by TagMF were found to be marginal. Despite the method's design to enhance the transparency of recommendations by integrating the semantics of tags, it may not be as effective as expected in making the recommendation process more understandable to users. This could potentially affect user satisfaction and trust in the system.

The potential for over-specialization is another concern. While we acknowledge that purely content-based approaches tend to over-specialize, recommending similar items, TagMF could also fall into this trap if not properly balanced. Over-specialization could limit the diversity of recommendations even if weights have been applied to tags, potentially leading to a filter bubble problem where users are only exposed to a narrow range of items that closely match their existing preferences.

Lastly, the effectiveness of TagMF relies heavily on user interaction. Users are expected to specify their interests via tags, which the system then uses to generate recommendations. If users do not interact with the system as expected, it could limit the effectiveness of the method. This reliance on user interaction could potentially limit the scalability in scenarios where user interaction is minimal or inconsistent.

2.3 Aspect-based Transparent Memories

In the previous section, we demonstrated how to enhance common recommendation techniques to impose transparency on latent information spaces. By exploring positive and negative associations between users, recommended items, and content attributes, we gained a better understanding of the derived semantics. In many cases, basing algorithmic recommendation decisions on such statistical correlations can contribute to a clearer comprehension of the recommended item. For instance, information about the tags associated with a movie can be useful in determining whether it aligns with personal preferences. However, deeper insights beyond statistical associations cannot be provided. For example, it is not possible to reliably determine why a specific sci-fi movie was preferred over many similar movies within the same genre.

It is, therefore, valuable to take a step back and contemplate the needs and expectations users might have for more comprehensive supportive evidence of recommendations. When individuals offer recommendations to each other in the real world, they often employ explanatory patterns that extend beyond simple item relationships with (dis-)liked attributes. Although justifications for a recommendation between friends may include such dimensions, further elaboration often involves more nuanced details about the specific aspects that make the suggestion appealing.

Moreover, the social nature of the individuals involved, along with their ability to communicate flexibly and adaptively, allows for closer alignment, fostering trust that is challenging to establish with a faceless entity like a technological recommender system [Kun*19]. Consequently, we believe that a stronger orientation toward human modes of reasoning to justify recommendations is reasonable. Our goal is to enhance the expressiveness of explanations to move beyond generic but prevalent schemes, such as "users who bought ... also bought ...", and provide richer cues about the suitability or unsuitability of an item. To achieve this, we draw inspiration from how people communicatively address explanatory concerns. Specifically, we aim to consider the expression of recommendations and their support through linguistic explanations as a communicative phenomenon that can be captured by principles of argumentation theory [WRM08; PM09; LT16].

However, since *explanations*⁵ for recommendations typically rely on statistical concepts rather than concrete system inferences, an explicit classification of this phenomenon within an argumentation-theoretical framework remains largely absent. In principle, though, supportive evidence generated during collaborative recommendations can be viewed as an application of the *argumentum ad populum* scheme. Fundamentally, an argumentation-based consideration of the connection between recommendation and explanation is both permissible and beneficial.

2.3.1 Recommendation & Argumentation

Essentially, we perceive a system-based recommendation as a distinct claim that the recipient is likely to find the suggested item particularly useful or interesting. However, unlike classical argumentation theory, a recommendation in this context does not claim general or exclusive validity; instead, it is personalized and may be based on local, temporal, or other contextual factors. Moreover, recommendations generally do not target a user's long-term preferences but

⁵See Section 2.3.4 for a critical discussion of the terminology employed in the context of explainable AI and recommendations.

instead aim to support decision-making in a specific interactive context, such as an online store, giving them a pronounced persuasive component.

From an argumentation theory perspective, it seems reasonable to support the original claim about an item’s suitability by reliably identifying meaningful and appropriate premises. To achieve this, we explore user-generated data sources often consulted in the absence of adequate system explanations. Specifically, we examine product reviews, which are widely available in many online contexts and serve as a rich source of experiences shared by presumably knowledgeable users. Due to their prevalence, they play a crucial guiding role in the decision-making process for undecided users [CM06; HLZ08; ZZ10]. Since they contain elaborate evaluations of item properties and are semi-structured, often following a certain argumentative flow, they represent a suitable data source for our purpose.

Extracting argument structures from such data sources is conducive to approximating the way people convey the pros and cons of an item. However, identifying argumentative patterns, whether manually or automatically, is far from straightforward. Human language is subject to numerous subtleties and ambiguities, making reliable extraction of argumentation graphs possible only in domains like jurisprudence, where the common structure is clear [PM09; MM11; XSA20]. In user reviews, which often feature slang and implicit or incomplete arguments, obtaining useful results is much more challenging [HG17]. For this reason, we will relax the strict constraints necessary for classifying arguments in the narrow sense. In principle, however, the presented method can be trained using only argumentative statements, as long as their prior identification can be performantly achieved in the future.

Even if arguments could be reliably detected, incorporating them as explainable components in a recommender system framework is not a simple task. One particular challenge lies in the relationship between user preferences and argument relevance. User opinions vary, and personal views on different aspects of a domain significantly contribute to its evaluation. For instance, it can be assumed that genre, story, visual aesthetics, and actors all influence whether a user wants to watch a movie. Furthermore, these individual aspects may conflict with each other. Effective persuasion, therefore, heavily depends on the target audience’s perspective on specific aspect categories. In this sense, the ability to identify arguments per se is only a partial solution to commissioning relevant premises. Instead, the recommender system should determine which pieces of available information are likely to be considered useful and how this information can be meaningfully presented in the context of a decision task. This scenario parallels the real world: when friends offer recommendations to each other, they carefully attach justifications to their claims that address their counterparts.

This implies two prerequisites for acquiring personally relevant arguments: First, the identification of domain aspects based on which arguments can be selected. Aspect identification can be described as a form of topic modeling in which the target entity, i.e., a domain item, is associated with certain attributes about which an opinion can be expressed. Second, the derivation of an assumption about how the target user evaluates these aspects. Thus, in addition to identifying *story* as a relevant aspect of the movie domain, the appropriateness of a recommendation largely depends on the ability to determine the types of stories a user is particularly interested in, i.e., the specific realization of the aspect. In order to achieve such a level of differentiation, we propose a method to establish preferential relations between users and these domain aspects via the route of textual feedback provided. In the present case, we have found that such personal inferences

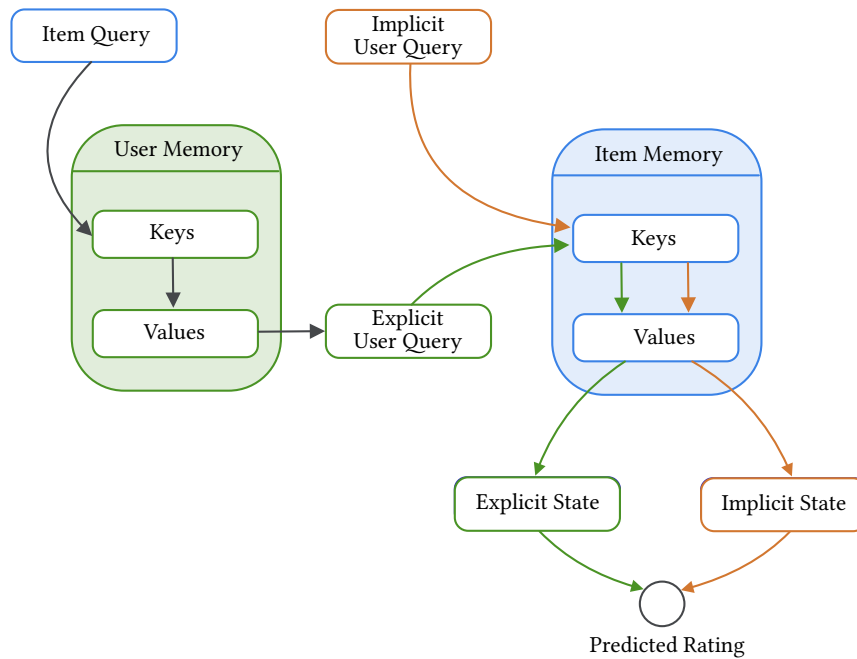


Figure 2.4 Simplified schematic illustration of the proposed rating prediction pipeline including read operations for the neural memory components.

are possible primarily on the basis of statements that contain polarization indicators, i.e., positive or negative sentiment.

2.3.2 Transparent Memories

Building on this idea, we introduce the Aspect-based Transparent Memories (ATM) approach, as detailed in [DKZ20], that imposes transparency on latent information spaces. This novel recommendation technique considers both positive and negative associations between recommended users, items, and content attributes to gain a deeper understanding of the derived semantics. While statistical correlations offer some insights, they have inherent limitations. Our approach addresses these by not only considering abstract behavioral feedback patterns as a basis for recommendations but also by elucidating the decision criteria individually expressed in reviews.

The primary objective of ATM is to associate a broad spectrum of content statements related to various domain aspects with users and items. This allows for the identification of specific aspects a user is interested in and the concrete statements that validate this interest. Moreover, our methodology can ascertain the degree to which an item embodies these inclinations, providing evidence through textual passages. To realize this, we employ advanced deep learning techniques, particularly neural memories [WCB15; Suk*15], adept at persistently encoding and decoding knowledge derived from textual content. The architecture we propose comprises two neural memories, each serving as arrays of slots for storing and retrieving aspect-oriented information (see Figure 2.4). One component encodes representations of statements from the target user, while the other, its counterpart, houses statements from other users about the target item.

Each memory is bifurcated into two subcomponents: aspect-based key vectors for addressing

operations and value vectors for content encoding. The model autonomously learns to pinpoint aspects, emphasizing the most pertinent ones in a textual unit. This extraction process embodies a form of neural attention, a cornerstone in contemporary deep learning architectures [BCB14; Vas*17]. By integrating these neural memories into established recommendation methods, our approach seeks to make recommendation computations more feasible by incorporating explicit statements on domain-specific aspects. This not only augments prediction accuracy but also clarifies the role of linguistically articulated preferences in recommendation selection.

To gauge the compatibility between a user and a target item, ATM engages both user and item memories in a mixed-initiative fashion. The explicit user state, rooted in textual content, and the implicit user representation, which discerns latent patterns, are merged to determine a potential match.

2.3.3 Aspect-based Interpretability

ATM aims to render otherwise unintelligible latent spaces interpretable, thus increasing transparency and accountability. We argue that such model-intrinsic justifications provide more faithful insights into the actual qualitative relationships between item features and recommendations, as opposed to *post-hoc explanations* [RSG16; STY17; MST20] which offer only approximations of internal model states. To establish a humanly understandable layer of transparency, we propose leveraging the attention values from ATM’s attentional mechanisms. These values can be used to generate two types of interpretability: First, representing the information space from a system perspective: By indicating which aspects the system pays attention to and clarifying how these aspects relate to concrete statements, the target user can evaluate how the system evaluates and structures input information. Second, extracting reasons for a concrete recommendation: Attention mechanisms can be used to identify statements of other users that support the claim of a recommendation.

Aspect Extraction To provide an overview of the information space, ATM can display details about the average distribution of aspects concerning the entire dataset or a specific item. The first step is to derive which aspects are primarily considered important by the system. A set of considered aspects can either be determined a priori or learned together with the other network parameters. The initial fixation can be achieved by assigning the aspect representation the word embedding of the respective aspect term or the average of embedding vectors of several terms, which together should represent a higher-order aspect. For example, in the case of *movie* recommendations, such a combination could consist of the embeddings of the words *story*, *story-telling*, *script*, etc. In contrast, automated identification of aspects can be achieved by extending the model’s cost function. Specifically, it is supplemented by an unsupervised loss that measures how well textual units can be reconstructed based solely on the combination of learned aspect embeddings. For a given textual unit, we can then determine the relative importance of each aspect.

Personal Aspect Importance Once salient aspects have been identified, ATM can use this information to further infer which aspects are particularly important for a specific target user. As before, this information can be obtained by averaging the aspect weights; only in this case, the textual target units should come solely from the corresponding user. However, merely displaying the distribution of the presumed personal aspect distribution provides insufficient transparency.

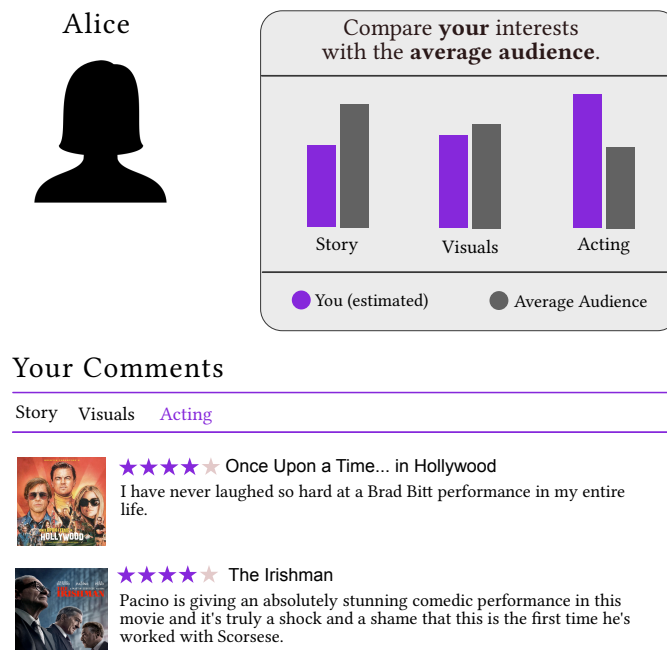
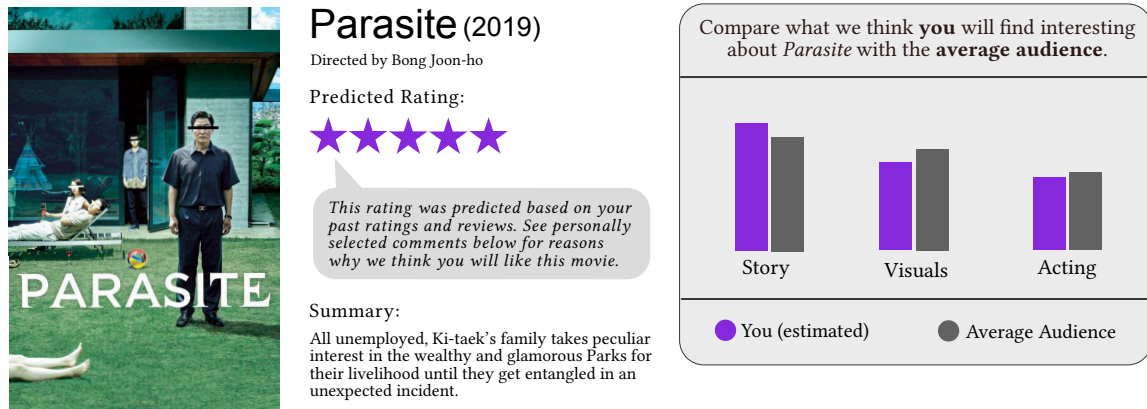


Figure 2.5 Exemplary user profile that depicting (assumed) personal and average importance for three aspects as well as the target user's comments sorted by aspects.

Instead, we can also display exemplary statements that have particularly contributed to the expression of a certain aspect weight. This allows the user to better assess whether they agree with the assessment of their assumed aspect preferences.

Recommendation Interpretation Another central interpretable component of ATM is a mechanism to communicate the logic behind a specific recommendation. ATM matches both the explicit and implicit user states with statements formulated by other users. The resulting attention weights then indicate which sentences have the greatest overlap with the vectorized user preferences. In other words, we assume that sentences with high weights are the best candidates to describe which properties of the target item the user is likely to like or dislike the most. Informally, the process can be exemplified as follows: Suppose a user has previously dealt extensively with storytelling in their reviews. More specifically, they seem to be interested in complex stories with a twist ending. ATM would then, based on concrete examples of such statements, create a user representation that semantically maps these preferences. If ATM now generates recommendations for this user, it would find a high correspondence between this preference representation and embeddings of statements that also deal with twist endings. As a result, not only would matching movies be identified in terms of the overall score, but individual statements containing specific information about the end of a movie would also be marked as salient. This applies equally to the implicit user representation.

In summary, ATM can be seen as the first step towards a fully developed argumentation-based recommender system. In its current state, it is primarily focused on extracting personally relevant information from reviews of other users. However, the selected content is not currently presented



Parasite (2019)
Directed by Bong Joon-ho

Predicted Rating:
★★★★★

This rating was predicted based on your past ratings and reviews. See personally selected comments below for reasons why we think you will like this movie.

Summary:
All unemployed, Ki-taek's family takes peculiar interest in the wealthy and glamorous Parks for their livelihood until they get entangled in an unexpected incident.

Compare what we think **you** will find interesting about *Parasite* with the **average audience**.

Aspect	You (estimated)	Average Audience
Story	High	Medium-High
Visuals	Medium	High
Acting	Low	Medium

● You (estimated) ● Average Audience

Comments

Story Visuals Acting Personalized

 Bob ★★★★★
Even as the plot ramps up and the tone starts to shift from dark comedy into tense thriller, Bong keeps a masterful hold on the reins.

Figure 2.6

Recommendation for the movie *Parasite* including the predicted rating, an overview of (assumed) aspect importance, and a personally selected comment that supports the predicted rating. Selection of comments can be personalized by toggling the respective radio button.

in a conventional argumentative form, as there is no structural knowledge about arguments represented in the model. This leads to some limitations, which we discuss in detail in the cited works. In particular, note that while the generated linguistic explanations can have a causal structure, the underlying attention mechanism is still based on purely correlational statistical processes.

2.3.4 Terminological Delineation

In the overall discourse of artificial intelligence and machine learning, the terminologies of *explanation*, *interpretability*, and *transparency* are often utilized interchangeably, yet they encapsulate subtly distinct connotations. An *explanation* typically denotes a comprehensible, simplified account of a model's output. Conversely, *interpretability* refers to the extent to which a human can comprehend the decision-making process of a model. *Transparency* pertains to the capacity to directly inspect a model's mechanisms.

Building on these concepts, it is important to explore the nuances that distinguish and sometimes intertwine explanation and *argumentation*. While the primary goal of argumentation is *persuasion*, explanation is principally concerned with *understanding*, trying to clarify how and why something is the way it is, without necessarily trying to convince. Since explanations often describe processes, mechanisms, or reasons behind a phenomenon, they do not necessarily follow the premise-conclusion structure of arguments. This distinction becomes particularly salient

when considering machine learning models. Argumentation in machine learning can be seen as a method to enhance interpretability by providing structured reasoning that traces a model's decision-making process. This rationale offers a deeper, more granular view than mere explanations, elucidating the series of logical steps or premises leading to a decision.

While ATM offers insights into its decision-making logic, its distinguishing feature is its grounding in argumentation theoretical principles. Rather than merely providing a linear account of its decisions, ATM constructs a set of supporting evidence to convince the end user of a particular suggestion, emphasizing its pronounced persuasive orientation. This is not to say that ATM is devoid of explanatory power. Certainly, users can offer concrete explanations in their reviews to justify their evaluation, and argumentative techniques can be used to support the validity of an explanation.

The term explanation can, however, potentially engender misconceptions in this context, as it may insinuate a level of understanding or causality that many machine learning models, including ATM, do not genuinely provide. This is particularly pertinent for post-hoc explainability methods like LIME [RSG16], which, for instance, strive to provide local explanations for individual predictions without necessarily encapsulating the global behavior of the model. These methods offer a simplified, often linear, approximation of the model's behavior proximate to a specific instance, which can be construed as a form of explanation. However, these explanations are not constructs of causality, but rather proxies for understanding the underlying decision-making process of the model.

In contrast, ATM, grounded in argumentation theoretical principles, is capable of offering structured premises to support its recommendations. This textual evidence, derived from user-generated data sources, provides a rationale for why a particular recommendation is made. By associating an array of content statements with users and items, ATM enables the determination of not only the aspects a user is ostensibly interested in, but also the identification of concrete statements that substantiate this assumption. This process mirrors human decision-making, where decisions are often justified using a series of reasons or evidence. This transcends the conventional notion of explanation in artificial intelligence, offering a form of interpretability that provides an avenue to understand the underlying decision-making process of the recommender system.

Moreover, ATM enhances the transparency of the recommender system by illuminating the specific aspects and statements that the system deems significant. This empowers users to comprehend the rationale behind recommendations and to evaluate how the system assesses and structures input information. However, it is crucial to note that while the extracted linguistic units can exhibit a causal structure, the underlying attention mechanism is predicated on purely correlational statistical processes. The resulting evidence thus expresses only an apparent causality, which is a critical distinction that necessitates explicit clarification. This phenomenon must be explored and elaborated in the future, for instance by referring to causal reasoning techniques [Sch22].

Given this distance between argumentation and explanation, it might be more accurate to characterize ATM as a method for *transparent interpretability*. This term amalgamates the notions of transparency and interpretability, underscoring the capacity of ATM to provide a lucid and comprehensible view of its interpretation of the underlying data landscape. It also circumvents

the potential misunderstanding that the explanations provided by ATM express a true causal relationship, when in reality they express an apparent causality.

In conclusion, while the term explanation is ubiquitously employed in the context of artificial intelligence and machine learning, it might not be entirely fitting for methods like ATM that aim to provide a more comprehensive understanding of the decision-making process of a model. Instead, terms like interpretability and transparency seem to be more appropriate, as they better encapsulate the ability of these methods to provide a lucid and comprehensible view of the underlying mechanisms of a model. As the field of artificial intelligence continues to evolve, it will be imperative to continue refining our terminology to accurately reflect the capabilities and limitations of different methods. This is particularly salient in the context of argumentation-related recommender systems like ATM, where the increase of transparency is closely tied to the identification and evaluation of meaningful and appropriate premises derived from user-generated data sources.

2.3.5 User Study

The evaluation of the ATM approach was conducted through an extensive user study with 136 participants aimed at assessing the quality of explanations as perceived by users. This evaluation was designed to identify the factors that contribute to the generation of good textual explanations in a decision-making task and how these factors interact with each other. The study was conducted in a between-subjects design, comparing the performance of ATM explanations against a state-of-the-art review-retrieval method.

The user study was designed to test the hypothesis that generating and presenting explanations in an aspect-based fashion would improve the subjective assessment of transparency compared to the retrieval of complete reviews. It was also hypothesized that the distillation of relevant sentences and grouping them by aspects would increase the overall textual quality of explanations.

The user study was conducted online with participants from Amazon Mechanical Turk, requiring participants to be located in the United States and to have an approval rate greater than 95%. Participants were assigned to two conditions in counter-balanced order in a between-subject design. In both conditions, participants were presented with five randomly sampled movies from the 100 most popular movies from the Amazon movie dataset. Each movie was depicted in terms of the title, runtime, director, actors, and a movie poster. Additionally, users received additional content extracted from user reviews subject to the condition they were assigned to. Users in the first condition were shown aspect-based excerpts identified by ATM, whereas users in the second condition received the most helpful reviews yielded by a review-retrieval method.

Statistical analyses based on *structural equation modeling* [UB12] revealed three central factors of textual quality: content adequacy, presentation adequacy, and linguistic adequacy. By tracing back mediation paths through the structural model, it was verified that ATM achieved significantly better results concerning overall quality. The greatest contributing factor to this finding was enhanced presentation adequacy. Hence, arranging textual explanations subject to salient aspects seemed to be a more insightful way of presenting information than merely retrieving complete reviews.

In summary, the user-centered evaluation of the ATM approach highlighted the importance of user studies in Recommender Systems and Explainable AI research. It revealed three central interacting factors contributing to the quality of supportive evidence. The results of the user study confirmed the effectiveness of ATM in identifying high-quality, aspect-based textual units that were well-received by users.

2.3.6 Limitations & Future Directions

ATM's reliance on user-generated reviews presents several challenges. The quality of these reviews can vary greatly, with some being superficial, off-topic, or even misleading. This variability can directly impact the quality of the explanations generated by the system. Moreover, the inherent ambiguity and subtlety of natural language, including linguistic complexities such as sarcasm, irony, and cultural references, can lead to misinterpretations and inaccuracies in the explanations generated. These issues highlight the need for future work to explore methods for assessing and controlling for review quality and to incorporate more sophisticated natural language processing techniques to better handle these complexities.

Identifying relevant aspects and mapping them to individual user preferences is another significant challenge. The diversity of user preferences and the complexity of certain aspects can pose difficulties, especially considering that the current approach does not account for preferential shifts over time. This suggests that dynamic aspect identification and personalization could be a valuable area for future research. However, ATM's approach is based on statistical correlations rather than explicit causal relationships. While the explanations generated can have a causal structure, the underlying attention mechanism is based on purely correlational statistical processes. This could potentially limit the depth and accuracy of the explanations, indicating a need for future work to explore the incorporation of causal reasoning techniques.

The computational complexity of ATM, which involves deep learning methods and attention mechanisms, could pose challenges in terms of computational efficiency and scalability, especially when dealing with large datasets. This highlights the need for future research to address these computational requirements. Simultaneously, there is a trade-off between transparency and complexity. While ATM aims to increase transparency, the complexity of the explanations generated might be overwhelming or confusing for some users. Balancing transparency with simplicity and understandability is a known challenge in explainable AI, and future work should continue to explore this balance.

Finally, the evaluation of the ATM approach was based on a user study with Amazon Mechanical Turk users from the United States, which may not be representative of all potential users of such a system. Furthermore, the study focused on the subjective assessment of explanation quality, which might not fully capture the effectiveness of the system in real-world scenarios. This suggests that future evaluations could benefit from a more diverse sample of users and a broader range of evaluation metrics.

2.4 Conclusion

In this chapter, we have delved into the realm of content-enhanced collaborative filtering systems, focusing on two novel approaches: Tag-enhanced Matrix Factorization (TagMF) and Aspect-

based Transparent Memories (ATM). Both methods aim to enhance the transparency and interpretability of recommender systems, thereby fostering a more user-centric approach.

TagMF leverages user-generated tags to augment the matrix factorization process, providing a more nuanced understanding of user-item interactions. This approach not only improves recommendation accuracy but also offers insights into the latent factors driving these recommendations. However, as we have discussed, TagMF is not without its limitations. The reliance on user-generated tags, while beneficial for capturing nuanced preferences, can also introduce noise and bias into the system. Furthermore, the approach assumes that tags are equally important across all users and items, which may not always hold true.

On the other hand, ATM takes a step further by incorporating argumentation theory principles into the recommendation process. By extracting and utilizing aspect-based information from user reviews, ATM provides a richer, more detailed context for recommendations. This approach allows for the generation of explanations that are more faithful to the actual relationships between item features and recommendations. Nevertheless, ATM also faces challenges, particularly in the extraction of reliable argumentative patterns from user reviews and the alignment of argument relevance with user preferences.

Looking forward, there are several avenues for further research and development. The exploration of new interaction mechanisms and the incorporation of additional data sources, such as social media activity or user browsing history, could enhance user control and personalization. Evaluating the impact of various content attributes on the recommendation process could provide valuable insights into their effectiveness for enhancing transparency and controllability. Furthermore, the development of hybrid recommendation models that combine different recommendation techniques could result in a more comprehensive and flexible recommender system.

In conclusion, the research presented in this chapter represents significant strides towards a more transparent, interpretable, and user-centric recommender system. By harnessing the power of content information and user-generated data, we have demonstrated the potential for greater user control, flexibility, and understanding in recommender systems. As we continue to refine these techniques and explore new approaches, we move closer to our goal of creating recommender systems that not only provide accurate recommendations but also enhance the user experience through increased transparency and interpretability.

Online Social Networks & Ideological Social Systems

In the previous chapter, we have delved into the technological foundations of modern recommender system architectures and discussed our approaches to elucidating the largely hidden data processing methods. In this context, we have introduced the concept of the *wisdom of the crowd* which serves as the fundamental mental model underlying prevalent collaborative filtering techniques. However, the term *crowd* suggests a level of uniformity that might obscure the actual functioning of recommendation algorithms. To be more specific, we have only touched upon the fact that the *crowd* might disagree about which target entities should be regarded as positive or negative. Depending on the domain and use case, there may well be substantial systematic differences between various user groups, potentially leading to the formation of preferential clusters at the topological level. Consequently, user populations often comprise a diverse or even antagonistic multitude of opinions.

As we pivot our focus towards online social networks, this discourse will lay the groundwork for a broader exploration of the societal ramifications of recommender systems. We will cast a critical eye on the prevailing evaluation methods in recommender systems research, advocating for a more holistic approach that acknowledges the influence of the recommender system itself on the creation of the database. In this light, we will underscore the relationship between the formation of individual filter bubbles and the potential emergence of preferential camps within the user population. This exploration will serve as a springboard for our subsequent, more detailed discussions on these topics.

Our ultimate goal in this chapter is to propose a simulation-based approach to understand and analyze social polarization. We aim to investigate the role of recommender systems in the formation and evolution of ideological groups within online social networks, and develop a method that considers the social dimensions in the analysis of recommender systems, thereby better assessing their far-reaching algorithmic influences.

3.1 The Dynamics of Recommender Systems & User Segregation

The question of how user groups become segregated is not as straightforward as it may seem. At first, it appears logical to assume that people have differing tastes and preferences, and therefore,

some users are more similar than others. After all, this is the very premise that personalization logic exploits. However, as soon as the recommender system actively intervenes in the relational process by proposing items, it enters into a mutual dependency relationship with the users. On one hand, user feedback supplies the system with behavioral surpluses, used to generate future recommendations. On the other hand, algorithmic content selection determines the items users can provide feedback on in the first place.

Research has demonstrated that this intricate dynamic can result in a narrowing of personal information horizons, a phenomenon popularized as *filter bubbles* [Par11]. This term essentially refers to a lack of diversification in recommended content, leading to a progressive degeneration of informational spaces and potentially hindering informed decision-making. This can reduce user satisfaction if the results appear too uniform and lacking novelty [Che*19; ZR21]. Consequently, recent academic efforts have sought to incorporate exploration concepts into recommendation calculations [KB16; ZR21].

Still, the subtle temporal and far-reaching dynamics emerging from the interaction between users and recommender systems remain enigmatic. We believe that the primary reasons for this lie in the excessive abstraction of concepts like *preference* or *need* often made during recommender system development as well as the lack of accessible analysis tools to pursue the investigation of long-term influences. As mentioned earlier, evaluation methods in this field focus on calculating objective metrics, such as prediction accuracy based on historical data. For example, the difference between an actual and predicted item score is commonly measured, or the degree to which an algorithmically derived item ranking matches a user's ranking is assessed. However, the influence of the recommender system itself on the creation of this database is usually disregarded.

Furthermore, this perspective on determining a system's quality is highly individualistic. As recommender systems typically aim to satisfy the needs of specific users, the social impact of information-deficient algorithmic distribution patterns remains largely unobserved, rendering the adoption of macroscopic views challenging. Nonetheless, we wish to emphasize the structural connection between the formation of individual filter bubbles and the potential emergence of preferential camps within the population [FGR16; FTA22]. Favoring connections among like-minded individuals ensures that they usually receive similar sets of recommendations, which can restrict information channels not only individually but also on a macroscopic scale.

In the case of entertainment products, which we have used as a running example in this thesis so far, the restriction of information horizons may lead to personally annoying consequences and be detrimental to individual user satisfaction. However, from a societal point of view, the formation of camps comprising blockbuster enthusiasts and art-house connoisseurs is not particularly relevant. Yet, the progressive divergence of user groups becomes problematic when it affects topics relevant to the general public.

The technologies we are discussing in this thesis not only play a role in generating recommendations for consumer products but also form the technological basis for online social networks [Twi23]. Unlike other applications that focus on connecting users and items, social networks also establish explicit and implicit connections between users, either directly or through message mediation. This results in a complex interplay between a platform's social fabric and the way recommendations are generated. For one, recommendation technology moderates which

actors meet on the platform at which time, significantly controlling communication processes. Conversely, newly formed follower and friendship relationships are incorporated into future recommendation calculations. Recommender systems consequently become full-fledged actors in digital social practice, exerting a significant influence on interpersonal phenomena.

Online social networks today represent an elementary channel for exchanging information about current, socially relevant topics. However, there is often little agreement on how to interpret ambivalent issues. Depending on social conditioning, cultural influences, or political attitudes, the same facts may be classified in very different ways. This form of differentiation already plays a central role in real-world social organization. Considering that the personalization logic of recommender systems can significantly contribute to potential camp formation in the user population, their role acquires its own significance.

From this perspective, specific digital phenomena, such as the emergence of so-called *echo chambers* [TBB21], can be described as the product of a complex interplay of social and technological influences. Echo chambers are generally described as environments where individuals are exposed primarily to information that aligns with and reinforces their existing beliefs, while alternative viewpoints are excluded or discredited. This often results in a self-perpetuating cycle that amplifies and entrenches opinions, leading to increased polarization and ideological isolation.

In summary, the importance of recommender systems in social networks extends beyond providing personally relevant content. They act as communication-moderating entities with potentially far-reaching societal influence. However, as previously explained, developing productive systems is primarily driven by maximizing individual engagement. In this work, instead, we aim to present a method designed to consider social dimensions when analyzing recommender systems.

Specifically, we investigate the role of recommender systems in the formation and evolution of ideological groups within online social networks, and develop a simulation method that considers the social dimensions in the analysis of recommender systems, thereby better assessing their far-reaching algorithmic influences. In doing so, we build on the insights we have already articulated in previous chapters, i.e., exploiting the properties of latent information spaces and their implicitly learned semantics. Our method proposes a novel means to modeling opinion dynamics by employing agent-based modeling [EA96; RG19] in unison with semantic latent spaces learned by machine learning, specifically knowledge-graph embeddings [Bor*13; Wan*17; Ham*18]. This approach allows for a more sophisticated representation of user opinions and preferences compared to existing methods based on one-dimensional attitude spaces [HK02]. Furthermore, our method considers the distribution of messages among users, including a recommendation component utilizing aforementioned machine learning technology. This combination enables the examination of the complex interaction between recommender systems and the establishment and spread of ideological fragments within online social networks.

As the aspect of social group formation has played a subordinate role in the context of recommender systems, we first adopt a social science perspective. We derive the importance of group-dynamic differentiation in social identity formation and its individual and societal effects, showing that social antagonisms are an inherent mode of operation in any social practice. We

then use these insights to describe echo chambers – or, more generally, ideologically shaped social groupings emanating from online contexts – as specific forms of social systems [Luh95] that primarily form their identities based on differentiation from opposing groups.

While a systems theory perspective allows us to understand user groups as social systems, it, importantly, also enables us to classify the role of recommender systems in connecting entities. In terms of systems theory, we can accurately describe when technological influencing factors become salient in system reproduction. Specifically, we assume that three entities enter into interdependent relationships: individual users (*psychic systems*), who depend on and influence social systems through transacted communication; the psychic systems and the *recommender system*, which reciprocally depend on each other; and the recommender system that moderates the mediation between social actors, significantly controlling the reproduction of *social systems*.

3.2 Group Identity & Differentiation

The cultivation of a self-referential identity through differentiation from other social groups is an essential feature of social organization. Anthropological studies suggest that even in early history, humans deliberately suppressed cultural diffusion when it served their own tribal hygiene [Mau00]. It is now clear, for example, that in North America there was a lively exchange between different tribes across the borders of supposed cultural areas. The people living there were, by and large, aware of what their neighbors were up to, and they also knew about foreign customs, arts, and technologies. Yet certain cultural traits, despite their potential usefulness, did not spread beyond tribal boundaries. Why is this the case?

Anthropologist Marcel Mauss's answer is that this is precisely what distinguishes cultures. As an example, he cites the fact that the Athabascans in Alaska stubbornly refused to adopt the kayaks of the Inuit, even though they were clearly better suited to the environment than their own boats [Mau00]. The Inuit, in turn, refused to use the Athabascan snowshoes. In anthropology, this feature of social organization is treated under the concept of *complementary schismogenesis*: "Societies live by borrowing from each other, but they define themselves rather by the refusal of borrowing than by its acceptance." [Mau00]. Basically then, Mauss concludes that people developed anything like a group identity at all only by comparing themselves to their neighbors [GW21]. Schismogenesis describes how societies in contact with each other end up in a dynamic of misunderstanding as they try to distinguish themselves from each other. In the classic example – the ancient poleis of Athens and Sparta in the 5th century B.C. – Marshall Sahlins notes:

"Dynamically interconnected, they were then reciprocally constituted [...] Athens was to Sparta as sea to land, cosmopolitan to xenophobic, commercial to autarkic, luxurious to frugal, democratic to oligarchic, urban to villageois, autochthonous to immigrant, logomaniac to laconic: one cannot finish enumerating the dichotomies [...] Athens and Sparta were antitypes." [Sah13]

Today, the psychological mechanisms underlying schismogenetic effects are well studied [Bor03]. Complementary schismogenesis describes essentially a process of differentiation of individual behavioral norms resulting from cumulative interaction between people with different cultural backgrounds. This implies striving for dominance on the one hand and submission on the other. According to Gregory Bateson's original definition, this kind of schismogenesis is a class struggle

between groups of people in which forms of behavior by one group provoke a corresponding counter-reaction by the other [Bat58]. The resulting reaction patterns are often characterized by a dominance-submission relationship and, by means of progressive, reciprocal amplification, harbor the potential to produce conflicts.

3.2.1 In-group & Out-group

The study of the dynamics of such group effects has been a major focus of social psychological research in recent decades. It is widely accepted that the underlying mechanisms by no means only apply to entire societies, but can contribute to the differentiation of diverse social groups of different sizes. People tend to construct their social identity through categorical schemes that relate to themselves and others. Thus, they perceive one and the same action very differently depending on whether they identify an actor as a member of their own group or a member of a foreign group [She*61; Taj74; Taj*79]. As a result, in-group actions are generally evaluated more favorably than ones assigned to an out-group. This discrepancy manifests itself even when people are randomly assigned to a group for no objective reason [cf. minimal group paradigm; Taj70]. Identification with and differentiation from social groups is apparently such a central component of social identity formation that even arbitrary assignment to a group is sufficient to form corresponding normative cognitive biases.

This fact is supported by recent neuroscience research, which provides evidence that the use of categorical schemas appears to be a fundamental mode of operation of the human brain [Mol13; CVB14; SD16; Sap17]. Consequently, categorizing people into social groups increases the perception that group members tend to be more alike. The out-group is perceived as a homogeneous mass, while the in-group is described as diverse [VBPC08]. This is especially true when negative characteristics are assigned.

Basically, then, psychological group effects are a necessity in dealing with reality, so that people can simplify complex phenomena and thus become capable of action. This reductionist mode of human world orientation manifests itself in more or less well internalized interpretation schemes and patterns, with the help of which people, partly consciously and partly unconsciously, structure reality both cognitively and normatively for their actions.

It should be noted, however, that the precise allocation of in- and out-groups may be subject to social contingencies [Sap17]. In some cases, membership may be clearly identifiable, for example through explicit affiliations such as citizenship. Often, however, boundaries are fluid or elude temporal addressability. For instance, established relationships between people in social networks can quickly disintegrate, while others are formed anew, leading to ever-new constellations. From the perspective of cognitive science, the categorization of people into groups describes a process that in many situations is based on prototypes, i.e., on a quantitative gradation of the membership of entities in certain categories [Ros73]. Accordingly, prototypes are central normative category elements from which the other elements can only be determined relatively with a greater or lesser *distance*. Thus, prototype semantics assumes that ideal categorizations never take place exactly, but can only gradually approximate an abstract prototype. Moreover, the intensity of the exercise of group-dynamic behavior is located on a fluid spectrum, ranging from mild to complete dehumanization of the *other* group. We will return to both aspects when we describe the technical implementation of our model.

3.2.2 Friend-Foe Dichotomy

First, however, we are interested in how psychological group effects can lead to a problematization of public opinion formation. According to Laclau and Mouffe, the antagonistic aspect underlying group identity formation is a constitutive feature of the political process [LM14]. For them, a harmonious society is impossible because the systematic distinction between *friend* and *foe* determines to an existential degree the functioning of a society.⁶ Accordingly, the tensions emerging from the constant negotiation of ideological differences and the struggle for hegemonic dominance are ultimately "inherent in every political practice and, strictly speaking, every social practice." [LM14] This structural antagonism, thus, extends far beyond the sphere of the political enterprise in the narrow sense; rather, it is the fundamental guiding difference of any ideologically shaped social form.⁷ According to [Mou99], the irrevocability of the antagonism of political actors is the central deficit of liberal theories of democracy, which assume that political antagonisms will disappear through greater individualization and rational discourse. Rather, the contingency of the process resulting from the constant differentiation of friend and foe ensures that the negotiation of power relations oscillates between the two extremes of their dissolution (total concentration and total equilibrium), but never reaches either state. As a result, the articulation of ideological boundaries is constantly recreated and renegotiated; there is no end point.

As a drastic example of group-dynamic friend-foe differentiation, the propaganda apparatus in Nazi Germany – especially with regard to anti-Semitism – is often used in social science studies [Sta85; AH97; NE02; Fro21]. Nazi propaganda can be understood as a mechanistic application of group psychology perverted to the extreme. Theodor W. Adorno describes Nazi ideology in this sense as a rigid structural unity based on a conscious and at the same time unconscious overall conception that "determines every word that is spoken" [Ado70, translated].

The obvious question that arises is how people can give up their own individuality and "voluntarily" submit to the coercive order of an authoritarian system. Without ignoring the complexity of the overall context, we would like to draw attention to the notion that the society of the Weimar Republic was essentially characterized by the fact that people at that time found themselves to a large extent in a state of social isolation and disempowerment [Are73]. The reasons for this are manifold. For our purposes at this point, it is important to note that the perceived or actual isolation was based on the loss of the immediate and human character of social interaction. People no longer perceived each other as autonomous agents capable of meaningful exchange, but rather as reified objects in terms of capitalist economic logic. Instead of encountering each other in empathic and considerate devotion, people "manipulated each other and treated each other as means to an end" [Ado70, translated]. Individualism and the naturalization of the laws of the market permeated all personal and social relations in the Weimar Republic to such an extent that today it is obvious that people had to be indifferent to each other if they perceived themselves exclusively as competitors.

In his influential work *The Crowd: A Study of the Popular Mind*, Gustave Le Bon described in 1895, long before the danger of German fascism became acute, how such a society can degenerate into a state of complete atomization. Le Bon characterizes a mass as de-individualized, without reason,

⁶Laclau and Mouffe explicitly refer to Carl Schmitt's *friend-enemy* distinction [Sch08] in their notion of the irrevocable opposition of political actors.

⁷The political thus does not constitute a specific space of social relations; "it is neither a center nor a node (the state), but a comprehensive horizon" [LM14].

easily led, and ready for violent action [LB02]. It may appear paradoxical at first to describe people as de-individualized in a society that explicitly promotes individualism. However, when we consider that the moment the project of individual self-realization fails (for whatever reason), there is nothing left but a lack of support in an increasingly complex social structure of atomized individuals, this argument takes on its own internal logic.

This lack of orientation in the face of the complexity of the world has always been a gateway for authoritarian ideologies. The Nazi *Volksgemeinschaft*, interestingly enough, corresponds exactly to Sigmund Freud's definition of a mass as a number of individuals who have substituted one and the same object for their ego-ideal and, as a result, have identified with each other in their ego [Fre89]. Consequently, followers of an authoritarian movement become completely absorbed into the group by sacrificing their own individuality.

According to Freud, the danger of such a massification of society arises precisely from the need to generate meaning. Natural social communities are replaced by virtual concepts of *Volksgemeinschaften*, which are characterized by the fact that they set themselves against others through the generation and dissemination of generalized antagonisms. One example of this is the tendency to *devalue the weak*, which can manifest itself catastrophically in the persecution of defenseless and weak minorities. Moreover, a pronounced characteristic of such uprooted people is that they develop aversion or even hatred for those who are *out*. In this sense, Freud's theory sheds light on the rigid distinction between the beloved in-group and the rejected out-group.

Freud's analyses, and later those of the Frankfurt School, are, however, by no means limited to the extreme cases of fascist agitation and radicalization, but describe a generally observable phenomenon of social group formation: "It is remarkable that in modern times this way of thinking and acting has apparently become so self-evident that it is seldom asked seriously enough why people love what is their own kind and hate what is different" [Fre89]. Freud believed that the dichotomy between one's own group and the group of others is so deeply rooted in the soul that it determines even those masses whose ideas are supposed to preclude such reactions:

"If another mass bond takes the place of the religious one, as the socialist one now seems to be succeeding, the same intolerance of outsiders will result as in the age of religious struggles, and if the differences of scientific views could ever acquire a similar significance for the masses, the same result would repeat itself as this motivation."
[Fre89, translated]

Consequently, Freud did not anticipate that the erosion of religious influence would necessarily dissolve the division between believers and unbelievers. Instead, he suggested that as traditional religious beliefs diminished, individuals might seek out other systems of belief or thought that satisfy similar psychological needs. This process could reinforce and even reify the distinction between those who believe and those who do not, albeit in new forms. Ironically, in the absence of the persuasive power of traditional religious doctrine, these new belief structures might be defended even more tenaciously, becoming structures in themselves that exist independently of their original ideological content.

Indeed, these emerging ideologies can also lead to new forms of group identity and cohesion. While supposed outsiders are met with hostile severity, intolerance disappears temporarily or permanently in the masses themselves through the psychological principles of mass formation. Within the group, individuals behave as if they were uniform, tolerating the peculiarities of the

other, putting themselves on an equal footing with them, and feeling no repulsion toward them. This phenomenon, which Adorno calls the *unity trick*, allows authoritarian agitators to maximize the diversity of the excluded while minimizing the perceived diversity within their own group [Ado70].

In summary, from a psychoanalytic point of view, we are dealing with a psychological impoverishment of the subject, who has surrendered to the object and replaced with it its most important component, the ego-ideal. In the individual's reference to the ego ideal, the ideology of the group is structurally manifested. Nazi ideology paradigmatically demonstrates how individuals experiencing a loss of control are susceptible to falling into a mode of social exclusion and marginalization when identitarian meaning is generated by a shared conspiracy myth such as Nazi anti-Semitism.

3.2.3 Delineation of Echo Chambers

In psychological terms, conspiracy narratives serve the purpose of assigning consistent meaning to a perceived overwhelming environment. The processes of meaning generation involve automatically triggered mechanisms of the cognitive system that are evolutionarily adaptive and develop dynamically over the course of socialization [SV08; Dou*19]. They are fundamental to human functioning in a complex environment. However, the search for meaning can have negative consequences if the reduction of complexity leads to a problematic distortion of perception. The tendency to believe in conspiracy myths can be interpreted as an exaggeration of the meaning-making process by assigning meaning to otherwise insignificant events. Studies have shown that feelings of insecurity or perceived loss of control increase the likelihood of engaging in superstitious thinking or believing in a conspiracy narrative. [GC17] showed in a study that perceptions of social isolation are positively correlated with both the need to generate meaning and belief in conspiracy narratives. In a subsequent analysis, they showed that generation of meaning significantly mediates the relationship between social isolation and conspiracy beliefs.

The dynamics of radicalization that are now thought to lie behind echo chambers incorporate precisely these principles of meaning creation and are based on the construction of worldviews that are deliberately diametrically opposed to the prevailing opinion. In this sense, the integration of conspiracy myths fulfills the function of demarcation and identity formation at the same time. On the one hand, an image of the enemy (*the elite*) defines an outside that is held responsible for a perceived loss of control. Such a lapse of trust in social institutions manifests itself, for example, in the discrediting of the credibility of certain authorities. On the other hand, dogmatic rejection can become an element of identity and community, insofar as people meet who share this *outside* perspective. According to the psychoanalytic view of a mass, we are thus dealing with psychological mechanisms that are closely related to the considerations of the Weimar Republic. In a sense, we are again facing an atomization of parts of society in many democracies today. In eastern Germany, for example, which is strongly characterized by rural exodus and structural poverty, the proportion of those who vote for (extreme) right-wing parties, fall for disinformation, or believe in conspiracy myths is significantly higher than in western Germany [Dec*22].

The genesis of echo chambers is hence based on a systematic alienation of their followers from generally accepted facts. Instead, members place themselves in an information structure that is

usually authoritarian and marks a clear distinction between friend and foe.

But the structural social establishment of conspiracy myths, which may even manifest itself as an organized form of violent protest, is only one extreme form of social differentiation. Cognitive distortions of meaning-making processes do not only affect socially isolated individuals. In reality, it is rarely the case that one side is completely right when the other is wrong. Political ideology and in-/out-group dynamics sometimes produce remarkable cognitive distortions that can lead to far-reaching social polarization effects. For example, Republicans and Democrats have systematically misjudged the risk of becoming seriously ill from coronavirus infection, but interestingly, in opposite directions [Leo21]. While Republicans often underestimated the danger, Democrats did the opposite: they overestimated the threat. The study shows that political position is the single most influential predictor of discrepancies. Thus, belief in the truth of certain facts appears to be explicitly used to differentiate themselves from each other, which is in line with the theory of complementary schismogenesis and the psychological group effects discussed earlier.

3.2.4 Derivation of Ideological Social Systems

Ideological disruptive factors and errors resulting from them can lead to the formation of deficient patterns of cognition. The epistemological study of ideology is concerned with the question of what desires, needs, and interests underlie these assumptions [Tep12]. Accordingly, the concept of ideology used here goes beyond a mere system of ideas and values or a specific world view. From an epistemological perspective, ideology describes epistemically deficient – i.e., distorted or illusory – thinking that is formed by reference to a particular system of ideas and values. It is thus a cognitive phenomenon of need-based thinking that produces misconceptions influenced by interests, needs, and desires. Here we assume universality: humans always operate within a belief system because this solves the basic orientation problem associated with the human condition. Thus, humans are worldview-bound creatures, even if they are not clearly aware of the premises of their frame of reference. Accordingly, it is possible for people to believe something to be true that is empirically inapplicable because the assumption in question meets their needs.

An ideology can thus be regarded as a world of ideas shared by a group or as an individual quantity. Ideologies are always formed in social frames of reference, but the bearers of the resulting structures are individual people. This means that not only explicitly formulated "large-scale ideologies with far-reaching group integration" [Tep12] are to be recognized as ideologies, but in general social systems⁸, which are structurally coupled to the cognitive process of their members through the mediation of ideological fragments or signs.

According to Francis Bacon's *doctrine of idols*, human reason is blinded by the fact that unconscious, value-based preliminary decisions play a fatal role in the conceptual fixation of a fact [BM57]. Thus, the medium of language used becomes a source of error. The handed down ideas and opinions have a certain authoritative character that can only be overcome by developing a critical faculty. If this does not happen, but the desire for certainty prevails, a hypothetical assumption about a context of meaning becomes an illusory certainty. A higher authority of a religious or areligious kind is postulated, to which absolute validity is attributed.

⁸What exactly we mean by systems, especially social systems, will be discussed in Section 3.3.

In light of this, any counter-values or positions must face rigorous negation. As such, when grappling with value system issues, individuals inevitably adopt a bipolar or dichotomous interpretation scheme. As [Tep12, translated] puts it, "The fundamental opponent or enemy is whoever wants the opposite of what one wants, especially in ideological or sociopolitical terms – even if one is not clearly aware of it." This dichotomy extends to the interpretation of facts, which vary depending on whether they concern the in-group or the out-group. [TS72, translated] assert that "emotionally charged black-and-white drawings and antithetical formulations are fundamentally necessary in the ideological and sociopolitical dimensions and in the realization of design goals in general."

This dichotomous perspective takes on heightened significance amidst the crises and rapid technological changes characterizing the contemporary era. Global events, such as climate change, a recent pandemic, a new war in Europe, or the evolving world order, have a profound, unsettling impact on societies, leaving many individuals disoriented. Concurrently, the swift progress of machine learning technologies – like large language models [Rad*19; Bro*20] or generative image generation methods [Kar*17; Goo*20] – adds to this sense of insecurity, particularly, but not exclusively, in online contexts. The rise of performant, stable diffusion models [Rom*22] makes machine-generated images nearly indistinguishable from real photographs. While assessing the authenticity of online content has always been challenging due to internet anonymity and source verification difficulties, the quality leap in these new models has precipitated a state where reliable authenticity assessments seem nearly unachievable. It appears that the age of unequivocal truth, if it ever existed, has come to an end. Looking ahead, as evidenced by the Republican reaction to Joe Biden's announcement of another presidential run [Gar23], it seems likely that entire election campaigns will be waged on the battleground of artificially generated images.

In the face of such disquieting transformations, individuals may turn to comforting ideologies to quell the cognitive dissonance they experience. For instance, even when new scientific evidence emerges about a viral mutation's threat level or lack thereof, such information may not be readily accepted. This resistance is especially likely if individuals have already integrated a network of ideologemes—fundamental units of ideological thought—into their broader worldview. Consequently, these individuals may ideologically rationalize their refusal to accept new information, externalizing it as an abstract concept. This process involves relinquishing some decision-making power as a means of avoiding responsibility, an idea that echoes Freud's concept of the shared ego-ideal.

3.3 Ideological Social Systems

Simulation-based modeling of social polarization often employs actor-theoretical methods, focusing on the individual – the *actor*. These methods model escalating polarization trends through self-referential operations [DeG74; Def*00; Wei*02; JA05; DGL13; Cha*14; DGM14; DV*15; PMC15; SB17; MBP18; MMT18; Tö18; Bau*20; PM20; Zha*20; SLL21; Sas*21; Arr*22; LRT22; Yua*22]. However, as discussed earlier, ideological directive structures significantly influence the cognitive processes underlying communicative behavior in social networks. These structures stem from socially acquired cognitions, often held as absolute truths and shared extensively among certain communication participants. In this context, an ideologically driven group presents an emergent self-referentiality that transcends the individual-centric focus of reductionist models.

Furthermore, such a group reflexively distinguishes itself from its social surroundings. We propose that it hence can be viewed as an operationally closed system, exhibiting unique structures to ensure its continuity. From this viewpoint, actor-theoretical models might encounter difficulties, as they are primarily designed to handle micro-sociological phenomena. This is not to say that actor-theoretical approaches are universally flawed, but rather that they might be limited in their ability to fully capture the dynamics of social polarization, which involves group-level processes and emergent phenomena. In the following, we aim to examine polarized group formation at a higher-order level, expanding the self-referential operation of actors to incorporate elements of social system formation.

3.3.1 Systems-theoretical Classification

Transferred to social systems theory, a social group generates its self-reference by demarcating itself from its environment, especially from other social systems. A social system, such as an ancient Greek polis, an authoritarian regime, or a conspiratorial ideological movement, generates its self-reference by performing two existential tasks [Luh95]: Firstly, it marks its boundary by means of operations, emphasizing the difference between system and environment. Subsequently, it generates and sustains itself internally by continuously linking operations in a manner that recursively refers back to previous system states. The system-environment difference thus represents the paradigmatic difference of any self-referential system:

”Self-reference can be realized in the actual operations of a system only when a self (whether as element, process, or system) can be identified through itself and set off as different from others. Systems must cope with the difference between identity and difference when they reproduce themselves as self-referential systems; in other words, reproduction is the management of this difference.” [Luh95]

Conversely, this implies that systems are structurally and intrinsically, and not just adaptively or sporadically, aligned to their environment, and their existence hinges on this relationship. Systems establish and sustain themselves by creating and upholding a difference from their environment. This demarcation serves as a regulatory mechanism for this distinction. Absent this differentiation, the very concept of self-reference would lose its significance, as the act of distinguishing oneself is foundational to self-referential processes: ”In this sense boundary maintenance is system maintenance” [Luh95].

3.3.1.1 System Boundaries

In social systems, unlike biological systems such as a cell, boundaries are initially ill-defined. They do not have membranes that allow for precise delineation. Instead, the difference between system and environment is mediated exclusively by meaning-constituted boundaries (*Sinnngrenzen*). A psychic system can still see its boundaries in its own body; social systems lack such cues. A meaning-constituted boundary, however, is not to be understood literally as an outer skin that acts like an organ. Rather, it relates the elements of a system to the system itself. Each element, seen in this way, makes an allocation decision and thus a boundary decision:

”Every communication in a social system, not just ones that cross the external boundaries, employs the system/environment difference and thereby contributes to determining or changing the system’s boundaries. Conversely, representations of bound-

aries serve to order the constitution of elements; they make it possible to assess which elements form in the system and which communications can be risked.” [Luh95]

Certain types of social systems, such as organizations, can clearly demarcate their system boundaries—membership serves as a clear boundary in this case. Similar identity-establishing mechanisms can be found in online social platforms. Here, participation is contingent on account registration, and communication is usually persistent, thereby becoming permanently observable. As such, at least on a mesoscopic level, differences between the system and its environment can be suitably defined. This makes it possible to view the entirety of communication that occurs on a platform as an operationally closed, self-referential social system. This system is structurally coupled with other systems in its environment. These could include real-world interaction systems (such as a conversation between friends leading to a tweet), other social media platforms (content from Instagram being shared on Twitter), or traditional media outlets (newspapers operating Twitter accounts)⁹.

The relationship between the general social system of mass media and the concrete media companies within it can be understood as a hierarchical and nested structure. Individual media companies, such as TV channels, newspapers, and digital platforms, function as sub-systems within the larger social system of mass media. These sub-systems influence each other and contribute to the overall functioning and dynamics of the mass media system. While each sub-system is operationally closed, they are structurally coupled with other sub-systems through exchanging information and structure formation, thereby shaping the overall media landscape. This structural coupling allows for the emergence of complex communication patterns and diverse media content. Understanding the interplay between the mass media system and its constituent sub-systems is crucial for analyzing how different media outlets influence public opinion, contribute to the creation of echo chambers, and shape the ideological landscape of society.

A more comprehensive classification of online social networks in the theoretical apparatus of social systems must be done elsewhere. Here, we are only interested in the differentiation of sub-systems from this overarching social system mediated by the network platform, which carry out their autopoiesis – their self-generated reproduction – through ideologically structured self- and foreign-reference.

3.3.1.2 Ideological Boundaries

Irrespective of the technical specifics of a network platform, the formation of social networks of any kind generally appears to be a highly precarious process, since it is initially subject to arbitrary addressability. The differentiation of ideological sub-systems is thus more difficult to grasp than the identity of the overarching supersystem, as their boundaries of meaning are considerably more volatile and can vary over time. Social networks are characterized by the fact that they have a broad distribution spectrum of communication at their disposal. Numerous actors

⁹The latter is an appealing example to illustrate why we assume closedness and self-referentiality in the case of platform networks, despite overlaps with their environment: The communication that takes place in the network is exclusively an element of its system, and only of it. The fact that *Tagesschau* posts on Twitter does not mean that it itself is part of the Twitter system. Just as Twitter does not become part of the classical distribution media in case tweets are received on television. The connection consists in a mode of structural coupling, which is made explicit, for example, by the fact that a newspaper editorial office operates an account on a platform.

are connected with each other through short distances – weak ties; completely separated communication networks are hardly to be observed in reality. However, the autopoiesis of a network system, if it succeeds, is supported by the fact that networks “produce a more or less specific range of services in the mode of reciprocal service communication” [Tac19, translated]. In the case of ideological systems in online social networks, this range of services corresponds roughly to maintaining a cognitive equilibrium by confirming one’s world view through agreement with like-minded people. This state of mental balance and stability, achieved through the alignment of an individual’s beliefs, attitudes, and values with their experiences, is maintained by engaging in reciprocal service communication within ideological systems. This process helps reduce cognitive dissonance [Fes62] by selectively exposing individuals to information and perspectives that confirm and validate their pre-existing beliefs, a phenomenon known as confirmation bias [Kla95]. However, while cognitive equilibria can contribute to a sense of psychological comfort and stability, they may also inhibit critical thinking and stifle the exploration of alternative perspectives, potentially perpetuating misinformation and contributing to social polarization.

In addition, ideological systems in online social networks serve a variety of functions that contribute to the reinforcement and perpetuation of shared beliefs among like-minded individuals. These functions include identity reinforcement, whereby individuals strengthen their sense of identity and belonging by associating themselves with a specific group or set of ideas; emotional support, which allows people to find validation and empathy for their beliefs, feelings, and experiences; information exchange, facilitating the sharing of resources and ideas to develop more nuanced perspectives on their beliefs; and mobilization and activism, providing a platform for collective action that empowers individuals to participate in social movements, political campaigns, or advocacy efforts.

Hence, ideological systems create and maintain their meaning-constituted boundaries through shared beliefs, values, and narratives. These boundaries not only help to define the system itself but also facilitate the differentiation between the system and its environment. For example, a political ideological system may be characterized by its adherence to certain principles, such as social equality, libertarianism, or nationalism. Members of the system, who share these beliefs, can identify themselves as part of the system while distinguishing themselves from those who hold different or opposing views. Ideological systems thus acquire a systemic character by turning to structural antagonisms. They generate a self-description qua formation of difference from the social environment (*us against them!*).

Although the boundaries of ideological systems in online social networks may appear porous and flexible, they still maintain their operational closure by using communication as a primary mechanism for self-reproduction. They may be formed and reinforced through the use of shared language, symbols, or narratives that resonate with the members of the ideological system. Hashtags, memes, or particular phrases can serve as markers that signal a system’s identity and differentiate it from its environment. For instance, supporters of a particular political ideology may use specific hashtags or linguistic codes to identify and connect with like-minded individuals, while distancing themselves from those who do not share their beliefs. These meaning-constituted boundaries, however, can also shift or evolve over time, as ideological systems adapt to new information, events, or changes in the broader social environment. This adaptability allows ideological systems to maintain their operational closure while remaining sensitive to their surroundings.

The system selectively processes information and communication, taking into account both internal and external inputs. In doing so, it continuously reaffirms its identity, incorporates new ideas or perspectives that align with its core values, and discards or neutralizes those that are incompatible. Ideological systems within online social networks can achieve self-reproduction under volatile and evolving connectivity by constantly adapting to the changing communication landscape. This is facilitated by the inherent flexibility of social networks, which allows them to reconfigure connections and communication patterns as needed. Thus, they achieve operational closure through their capacity to process, adapt, and integrate communication inputs selectively, while constantly reaffirming their core principles and adapting to the changing connectivity landscape.

We suggest that such ideological coherence forces specific modes of communication and therefore generates observable behavior in the form of connective communication (*Anschlusskommunikation*). Internal system communication is moderated in such a way as to increase the likelihood of a connectible operation. The autopoiesis of an ideological system takes place through the (re-)production of identitarian ideologems which is existentially activated by the communicative expectations that alter directs at ego. One identifies the other as a member of their own or of a foreign group on the basis of categorical schemata or previous encounters, thereby provoking the psychological phenomena of group behavior already discussed.

As indicated before, individuals often exhibit a tendency to seek out familiar sources and supportive evidence (see selective exposure or confirmation bias), a behavior that could be interpreted as a systemic response to the complexities of double contingency (*Doppelte Kontingenz*) as outlined by [Luh95]. This tendency reflects a broader theme in Luhmann's theory, where social systems and communication structures emerge to reduce the uncertainties inherent in interpersonal interactions and manifest themselves at the group level. In this scenario, the communication expectations of ego are shaped by, and in turn shape, the anticipated reactions of alter. These mutual expectations, through repeated interactions or shared experiences, might cultivate a sense of trust or comfort. The established trust or familiarity could engender a sense of predictability amidst the inherent uncertainties of interpersonal interactions. As individuals navigate the complexities of double contingency, the cultivated trust or familiarity becomes a heuristic for reducing uncertainty. Consequently, this drives a preference for familiar sources that are perceived to align with one's own perspectives, as these sources represent a continuation of the predictability engendered by the established trust or familiarity. This preference for familiar sources thus emerges as a systemic response aimed at maintaining or enhancing the connectability in communication, which is pivotal for the sustenance and evolution of the ideological system under discussion.

The resulting notion of boundary is thus an answer to the question of how, in the face of ongoing cross-border communication and constantly reproducing surpluses of meaning taking place on social platforms, it is nevertheless possible for system-internal connective communication to occur reliably. At the same time, however, this also establishes that the drawing of boundaries in the case of ideological systems is not based on a single mechanism (like membership in the case of organizations), "but rather on a proliferation of social (who?), factual (what?), and temporal (when?) constraints on network communication" [Tac19, translated]. Networks find support exclusively in themselves, precisely in the particularity of the connection and in the interlocking of the social, factual, and temporal structures, each of which applies only to them. In a sense,

this makes them both arbitrary and universal: they can occur anywhere in society and are at the same time ephemeral [Tac19].

The boundaries that define *us* and *them* are not static, but are dynamically negotiated through ongoing interactions with the surrounding ideological milieu. This distinction serves as a fundamental mode of operation, delineating the ideological system from the external environment and thereby rendering it identifiable in the midst of the digital ecosystem. This demarcation is not a mere abstraction, but a pragmatic mechanism that allows for the reduction of complexity within the ideological system. By maintaining a distinct external boundary, ideological systems aim to reduce the complexity of the multifaceted ideological milieu online. The boundary does not act merely as a sieve filtering out discordant narratives, but organizes them in a manner that bolsters the ideological system's autopoiesis. Through engaging with discordant narratives, the ideological system orchestrates interactions in a way that upholds its own ideological foundation, be it through disagreement with the counter narrative's content or discrediting its author. This encapsulation allows the in-group to navigate the ideological landscape with a sense of clarity and purpose, shielded from conflicting external ideologies yet actively engaging with them to reaffirm the ideological coherence of the system. For instance, within an echo chamber, discordant opinions might be presented and dissected to reinforce the in-group's shared narrative, thereby reducing the cognitive load on the in-group and making the internal narrative appear more cogent and compelling.

Moreover, the external boundary requires constant maintenance to uphold the integrity of the ideological system in a turbulent online environment. Luhmann's concept of re-entry is particularly illuminating here; the distinctions that define the system are continually reintroduced and reinforced to maintain the boundary. The friend-foe distinction is not a one-time declaration, but a repeated narrative in which the delineation is revisited and reasserted in response to evolving external ideological interactions. This incessant re-entry serves to reinforce the external boundary, ensuring that the ideological system remains distinct and coherent amidst the digital maelstrom.

Summarized, our hypothesis is that differentiation according to a friend-foe scheme can, under certain conditions, lead to a stabilization of intra-system communication and thus make it observable and determinable. Thus, messages from supposedly like-minded people are more likely to be passed on. Cross-border communication, on the other hand, is met with pre-emptive mistrust. Accordingly, while communication is still evaluated according to content or knowledge of the author, the primary mode of differentiation is realized with respect to the expectations of non-members who are shaped by a conflicting worldview. At the same time, members of an ideological group may engage in discussions, debates, or the sharing of information that aligns with their shared beliefs. This communication process allows the group to continuously reaffirm its identity and create a sense of cohesion among its members. Furthermore, the dynamics of online communication, such as liking, sharing, or reposting content, can create feedback loops that reinforce the self-referential nature of these systems. When group members are exposed to content that supports their beliefs, they are more likely to interact with it and share it with others in their network. This, in turn, increases the visibility of the content and further amplifies the group's shared perspective. These feedback loops can lead to the perpetuation of polarized group formation, as they encourage individuals to predominantly engage with content and people that confirm their existing beliefs, while isolating them from alternative viewpoints. The result is

a self-reinforcing cycle that strengthens the operational closure of the ideological system and contributes to the deepening of social polarization. On a topological level, this relationship is expressed by a densification of certain communication channels, so that – and this has been amply demonstrated by network analysis research [e.g. Les*09] – distinguishable communities can usually be identified. We argue that these communities can potentially have systemic character if they are based on a shared ideological structure.

Hence, it also becomes clear that usually numerous ideological systems stand in environmental relationships with each other. This significantly increases their internal complexity, which in turn requires a corresponding reduction of uncertainty. According to psychological theories of group behavior, this reduction takes place through generalization and then the formation of categorical schemas and stereotypes. If we assume that ideological systems actualize themselves out of the group-dynamic identification of a psychic system with it, then the influx of a message from *outside* contributes to the autopoiesis of the ideological system only if it recognizes the production of connective communication as conducive to its own continued existence. The distribution of messages from supposed outsiders would then be assigned to the autopoietic process. We will return to this when we describe the technical implementation of boundary setting.

3.3.2 Structural Coupling

As we have seen in the previous section, a self-referential system reproduces itself only if it exhibits connective operations. In social systems, connectivity is realized through the potential to generate connective communication. Thus, the basic process of social systems that produces system elements is communication. In this way, systems theory departs from the assumptions of actor theory [Mur98; CB07], according to which actions are the last ascribable individual choices. According to Luhmann, the social disappears if a communicative act is not followed by a subsequent action. He thus foregrounds the intersubjective character of the reproduction of social systems and excludes a psychological determination of the unity of the elements. Although Luhmann says that communication and action cannot be separated, they can be distinguished: "The elementary process constituting the social domain as a special reality is a process of communication. In order to steer itself, however, this process must be reduced to action, decomposed into actions." [Luh95].

In this context, systems theory is often accused of abstracting too much from observable actors by replacing action with communication. As [Bra97] poignantly summarizes, some authors reject this exclusion of actors and argue that humans, according to Maturana's theory of biological systems, are the essential element in ensuring autopoiesis. Beyerle criticizes Luhmann for not being able to explain how communication is generated: "He can make plausible the self-reference of sub-systems through communication, but since he combines society and communication, he cannot explain the emergence of this communication" [Bey94, translated]. In Maturana's view, the act of interaction in the action-theoretical sense makes humans an immediate and indispensable component of social systems. Consequently, society as an emergent system cannot be explained in terms of communication alone [Gö02].

Ulrich counters this interpretation by stating that the autopoiesis of social systems is simply no longer possible if humans are included as constituents. This is only guaranteed "if the same

elements connect on the basis of system-specific modes of operation” [Ulr94, translated]. Consciousness, however, is a system-independent element of social systems. Furthermore, the criticism that systems theory abstracts too much from humans can be countered by the fact that Luhmann by no means excludes the individual from the equation of social interrelations. Quite the contrary: Luhmann sees human beings as the precondition for everything else. What he means by this is that a human being is far too complex an entity to be integrated as a unit of analysis into the theory of social systems and the definition of society with the necessary selectivity.

To deal with the complexity of psychic systems, Luhmann abandons their comprehensive definition and instead considers the structural coupling between social and psychic systems. All communication is structurally coupled to consciousness: “Communication (in every operation) is totally dependent on consciousness – if only because only consciousness, not communication itself, can perceive sensually, and neither oral nor written communication could function without perceptual performances” [Luh97, translated]. Accordingly, communication systems, like consciousness systems, are operationally closed systems that are structurally oriented toward each other but cannot maintain direct contact. “Only one consciousness can think (but not: think into another consciousness), and only society can communicate. And in both cases they are self-operations of an operationally closed, structurally determined system” [Luh97, translated].

But the structural coupling is there, unnoticed and incessant: “It works even and especially when one does not think about it and does not speak about it – just as one can take the next step during a walk without thinking about the physically necessary own weight for it” [Luh97, translated]. The mutual intransparency of the coupled systems is thus a necessary condition for structural coupling, because otherwise the endogenous operations of the systems could not be synchronized. Through structural coupling, a system can be connected to highly complex environmental conditions without having to work out or reconstruct their complexity.

The concept of structural coupling bridges the gap between the sociological discussion of social systems and the development of our simulation framework to model online social polarization. By utilizing this notion, we can examine the intricate relationship between social systems, particularly ideological ones, and the decision-making processes involved in connective communication. Structural coupling allows us to better understand how different self-referential systems interact, offering valuable insights into the complex dynamics that underpin social polarization. As a result, this theoretical starting point lays the groundwork for the subsequent derivation of a simulation framework that builds upon these concepts to model the phenomenon of social polarization more effectively.

3.3.2.1 Structure Formation

In order to realize its own connectivity, a social system must form specific selection schemes that enable “recognition and repetition, thus condensing identities [...] and confirming them in ever new situations, thus generalizing them” [Luh97, translated]. Such structure formation then allows for the reduction of environmental disorder within the system through forms of remembering and forgetting. In an abstract sense, the concept of structure can be related to communication insofar as structures that link communication contain information and are thus *world structures*. Psychic and social systems thus stand in an interpenetrating relation to each other through the formation of structures. Penetration means first of all that one system makes its own complexity

available for the construction of another system. In this sense, social systems presuppose human beings. Accordingly, interpenetration exists when this state of affairs is reciprocal, "when both systems enable each other by introducing their own already-constituted complexity into each other" [Luh95]. In interpenetration, therefore, the receiving system influences the penetrating system through structure formation.

Communication thus affects the formation of linguistic and non-linguistic structures on the one hand, but is itself shaped by different structures on the other. Thus, when communication takes place between *ego* and *alter*, both partners direct an expectation of communication toward each other based on the structures established between them. Since *ego* and *alter* are mutually indeterminable and encounter each other as *black boxes*, they generate expectations for a follow-up communication that can be inferred on the one hand from previous communication and on the other hand from other structurally significant information. For example, when *ego* and *alter* communicate on a networking platform, *alter* may consult *ego*'s personal profile to increase their own predictive ability. Again, however, it is important to emphasize that structural referencing in psychic systems is largely unconscious.

The most important carrier of structures is language itself. According to the view of (linguistic) structuralism, language is a system made up of different signs, e.g., words, suffixes, idioms, etc. The meaning of a sign is defined exclusively by its relation to other signs in the system and explicitly not by a relation to an external reality [Sau11]. In this sense, the meaning of a linguistic unit emerges from its co-occurrence with other units in certain contexts. Thus, meaning is not derived from some intrinsic properties, such as the letters of which it is composed or the way the tones sound when it is pronounced. From this perspective, dictionary definitions do not capture the meaning of a word as well.

Saussure's groundbreaking work on the theory of linguistic structuralism revolutionized the way meaning is understood within language. Central to this notion is the distinction between the *signifier* – the sound-image or word – and the *signified* – the concept or idea associated with the word. Saussure posited that the relationship between the signifier and the signified is arbitrary, as there is no intrinsic connection between the word and the object or idea it represents. Furthermore, Saussure emphasized the principle of differentiability in the construction of meaning. This principle asserts that the meaning of a linguistic unit is determined by its relationship to other units within the same linguistic system, rather than by any intrinsic properties. Consequently, meaning is derived from the differences and oppositions between words, as opposed to their inherent qualities. In this context, Saussure also introduced the concept of binary oppositions, which are sets of contrasting terms, such as feminine/masculine or singular/plural, that help to define and create meaning within a linguistic system.

The structural interdependence between signs is both individually internalized during socialization (and thus determinable as a psychological phenomenon) and culturally generalized beyond the individual (and thus sociologically relevant). From a structuralist perspective, it is precisely this hypothesized cultural system that significantly determines our perception. Thus, ideological fragments manifest themselves especially on the linguistic level by differentiating secondary systems of meaning from the primary system of related signs. According to Roland Barthes, this is the mechanism by which myths – and, by analogy, ideological forms of meaning – can emerge

as signs of a higher order¹⁰ [Bar79].

Through the structural coupling of psychic and social (ideological) systems, such systems of meaning become relevant for communication and action. They decisively determine which communication expectations are ego and alter¹¹ direct at each other and thus influence both the probability and the form of connective communication.

3.3.2.2 Derivation of the Conceptual Framework

What is distinctive about social network platforms is that the communication process is largely observable and persistent. While in interaction systems that result from direct face-to-face encounters, communication is primarily verbal, the exchange of information on a platform is mostly done in writing or in platform-specific forms of communication such as the *Like*-button. Thus, while communication-relevant structures remain latent, they manifest themselves to some extent in the way the relationalization of elements takes place as a communication process. Our hypothesis is that by observing and analyzing network communication, it is possible to approximate the latent but significant structures that, on the one hand, determine the characteristics of the overall system and, on the other hand, allow the differentiation of specific (in our case: ideological) sub-systems.

Recently, the study of language and communication on social platforms has been increasingly pursued with the help of machine learning algorithms [OMK20]. These algorithms are often based on methods that project the relevant entities into a latent information space, a concept we have extensively dealt with previously (cf. Chapter 2). For example, if one wants to investigate which linguistic forms of meaning manifest themselves in the digital sphere on certain topics, the corresponding semantics can be derived from large data sets. *Embedding* vectors can be computed by considering which words often occur together in certain contexts. The derived positions of these vectors in the information space allows a comparison between the learned concepts.

The ability of learning algorithms and the latent information spaces they generate to map semantics has been well demonstrated. The best known example of this is the overlap between the intuitive semantics of a word and its learned position in the information space by the Word2Vec algorithm [Mik*13]. Walking between the position vectors *queen* and *king* in the semantic information space, the direction and length of the translation vector is almost identical to that between the concepts *woman* and *man*. This difference can be calculated by subtracting the corresponding embedding vectors, e.g. *woman* and *man*. The result of this subtraction is a vector that can be interpreted as a translation operation in the latent space and semantically represents the concepts *feminine* and *masculine* at their respective poles. In a figurative sense, then, relativity in the positioning of entities represents an approximation to the linguistic concept of the signified. In the latent information spaces that can be generated by machine learning technologies, structure and semantics thus coincide.

¹⁰From this perspective, a system-internal, ideologically shaped conceptual apparatus can be described as a symbolically generalized medium of communication that transforms the improbability of connective communication into probability and thus increases the likelihood of the autopoiesis of an ideological system.

¹¹Luhmann explicitly leaves open whether psychic or social systems are meant by the terms ego and alter. Accordingly, a psychic system (ego) can also direct a communication expectation to a social system (alter), e.g., to the entire population of users on a network platform or to the circle of people directly connected to them.

This gives rise to an appealing analogy between machine learning algorithms and the theory of (linguistic) structuralism: while the latter assumes that the meaning of linguistic units can be derived from their relationality with other elements of the system, the embedding of a word only acquires meaning when it is compared with those of other terms, for example by computing distances in the latent information space.

We now want to transfer this property to communication in social networks; however, not with regard to the semantics of the language used per se, but in the sense of the systems-theoretical communication process as a relationalization of communicative elements. Therefore, we base the embedding in an information space on the network topology in order to map and model the relations between psychic systems, their produced communication units (e.g. posts and reposts) and later on ideological social systems. Here we consider the communication units as cultural fragments that can contain meaning structures of primary and secondary order signs and thus (re-)produce ideologemes both in terms of content and in terms of forms of subsequent follow-up communication.

In summary, we assume that machine learning algorithms can extract latent structures from the network topology and the communication behavior of users. These learned structures approximate the latent structures of reality and are thus the starting point of our considerations. Based on this, we will then simulate both the communication behavior of psychic systems and, bottom-up, the genesis of ideological systems.

3.4 Conceptual Framework

Moving from this theoretical framework to the substantive components of the proposed simulation model requires an understanding of how these structural underpinnings can be utilized to model digital communication processes. The importance of latent information spaces in this endeavor cannot be overstated as they serve as multidimensional landscapes that encapsulate the complexities of social interactions and communication dynamics. Within these spaces, we will conceptualize a radius of acceptance that delineates the extent to which individuals are inclined to engage in communication, a concept that is integral to the functionality of the model. The uniqueness of the present approach lies in the integration of human communication behavior and recommendation logic, both of which are derived from the same latent space. This unification aims to provide a comprehensive model capable of capturing the nuanced interplay between social interactions and the influence of recommender systems on the emergence of polarization effects.

The latent space model we want to present in this thesis is essentially based on the assumption that generalizable structural relationships can be derived between the different nodes of a social graph – see [DZ21; DZ23] for a detailed derivation of the model. We do not merely consider social connections between actors, however, but intermediate their relationship through the messages exchanged over the network. The resulting heterogeneous communication graph \mathcal{G}_c is, hence, comprised of a set of user nodes \mathcal{V}_u and message nodes \mathcal{V}_m . Users are initially connected to messages via two types of edges: First, messages can be composed by them (τ_s), second, they can react to incoming messages in the form of connective communication (τ_r). This allows us, for example, to specifically capture which messages a user has reacted to in order to assess their personal behavioral feedback.

Our working hypothesis is that it is possible to model the psychological and sociological concepts by operationalizing the semantics of a latent model learned based on this structural information. Specifically, we want to exploit the derived topological correlations between entities to implement geometric operators together with the latent representation of graph nodes. Detailed explanations of such geometric operators can be found in [Ham*18; DZ21]. For the present case, we will restrict ourselves to a projection operator \mathcal{P} , which translates a given latent representation $\mathbf{q} \in \mathbb{R}^d$ in a space of dimensionality d subject to a certain edge type τ with representation $\mathbf{r}_\tau \in \mathbb{R}^d$:

$$\mathcal{P}(\mathbf{q}, \tau) = \mathbf{q} + \mathbf{r}_\tau \quad (3.1)$$

For instance, we may apply geometric operator $\mathcal{P}(\mathbf{u}, \tau_r)$ to user node $u \in \mathcal{V}_u$ to identify a region in information space where message nodes are in close proximity to each other that this user has a high probability of propagating. Note that \mathbf{q} can represent the original latent representation of a concrete graph node as well as one that has already been shifted one or more times. For example, $\mathcal{P}(\mathcal{P}(\mathbf{u}, \tau_r), \tau_s)$ searches for actual linked users as well as potentially linkable matches for a target user u .

A projected latent representation can subsequently be interpreted geometrically by examining the distances to graph nodes of a certain type. For this we introduce δ as a second operator which specifies the distance between a projected representation \mathbf{q} and a target node v :

$$\delta(\mathbf{q}, \mathbf{v}) = \frac{1}{2} \frac{\text{Var}(\mathbf{q} - \mathbf{v})}{\text{Var}(\mathbf{q}) + \text{Var}(\mathbf{v})}, \quad (3.2)$$

where Var is the variance of the input vector and $\mathbf{v} \in \mathbb{R}^d$ is the representation of v .

For the combined application of one or more projection operations to an anchor node v_1 with a final calculation of distance to a target node v_2 , we write $\delta_{\mathcal{P}}(v_1, v_2 \mid \mathbf{T})$, where $\mathbf{T} = (\tau_1, \dots, \tau_n)$ is a tuple of projection edges. Starting from this central operation, we now want to address the question of how we can exploit the semantics of the latent information space to relate the entities of a social graph and formulate a simulation procedure from it. Specifically, we want to take the geometric relationships between entities as a basis for an opinion dynamics model that determines whether or not a user will engage in connective communication.

3.4.1 Psychic & Social Systems

Suppose the situation of a user, faced with a novel input message, who needs to decide whether or not to react to it in some way, for instance by further propagating it. From a systems-theoretical perspective, the occurrence of this action describes the production of a connectable operation in the form of connective communication and thus the basic prerequisite for the autopoiesis of a social system. Accordingly, the sending of a message alone cannot be regarded as communication. Only when this message is followed by a communicative reaction is the reproduction of the system ensured. Ignoring ideological factors for a moment, after first turns to the overarching supersystem to generate an expectation of communication. Accordingly, they decide, on the

basis of the structures established between them and the other participants involved, whether to produce a follow-up communication. These selection schemes essentially refer to his or her previous experience and thus in particular to the concrete personal communication network, which is determined by (unilateral or reciprocal) social ties as well as the technological dissemination patterns derived from them.

It is still true that structures are mainly communicated and updated via linguistic cues, and that content-related criteria therefore play an elementary role in establishing a connective reaction. At the same time, however, further dissemination always takes place within a frame of social reference. Thus, alter not only evaluates the content of a message transmitted by ego, but also determines the extent to which they consider this content to be worthy of dissemination to their addressed audience.

We want to illustrate this connection by means of the described couplings in the latent information space formed by the edges of a communication graph. The underlying heuristics can be described as follows: The learning process of the applied machine learning technology results in a structural connection between the individual graph nodes. By applying a backpropagation procedure, the positions of both the head and the tail of a relation are adjusted depending on the measured error of the prediction. Thus, the geometric arrangement in information space is not unilaterally determined, but is subject to mutual dependencies. Beyond, this results in a complex network of transitive influences, so that, for instance, users with similar communication behaviour will also be topologically close to each other. Since alter's position also depends on those of the other communication participants, they are structurally coupled to the social environment. We speak here explicitly of structural coupling in the systems-theoretical sense because we assume that the derived relationships are an approximation of the actual structural coupling in the real world.

3.4.1.1 Individual Demarcation

Based on the principle of similarity, which is very much in line with the idea of the *wisdom of crowd*, we now want to make assumptions about the probability of connective communication. The sense-making processes that are triggered during the evaluation of new messages are initially attempts to classify what is perceived into a personally meaningful structure. This individual frame of reference implies that the generation of meaning is anchored in the construction of one's own identity [Wei95]. However, sense-making processes require quite a range of cognitive heuristics, mental shortcuts so to speak, to function efficiently. Important to our case is the human tendency to prefer things they are well acquainted with which implies that information, that contradicts one's internalized assumption about a phenomenon or that cannot easily be integrated into one's belief system, often tends to be disregarded or ignored [Fre86].

Translated to latent space semantics, we assume that the likelihood of connective communication is high if a new message is structurally similar to messages with which the user has interacted in the past. In order to transfer this characteristic to a model of opinion dynamics, we refer to advances in agent-based modeling (ABM) [JA05]. The core idea behind ABM is that the actions of certain agents, for instance humans, can be modeled by applying a mathematical decision function simulating a reaction towards incoming new input. Specifically, we take inspiration from social judgment theory [SH61] to model this with respect to a concept called *latitude of*

acceptance. Whenever a new bit of information comes to the attention of an individual, we apply a bounded confidence model to determine the likelihood of connecting communication as:

$$\mathcal{L}_1(\tau_r(u, m)) = \frac{\lambda^\mu}{\delta_\varphi(u, m | \tau_r)^\mu + \lambda^\mu}, \quad (3.3)$$

where $u \in \mathcal{V}_u$ and $m \in \mathcal{V}_m$ are user and message nodes respectively, $\lambda \in [0, 1]$ is the latitude of acceptance, and $\mu \in \mathbb{N}$ is a sharpness parameter that determines how steeply the likelihood of connecting communication drops from one to zero around the latitude of acceptance. Informally, a user representation is projected into the area in latent space where messages with already established edges are located. Then, the distance to the target message is calculated and fed into a non-linear decision function.

3.4.1.2 Ideological Demarcation

As we have previously described, we view the cultivation of identity through differentiation from other social groups as an essential feature of social organization in the digital sphere. At the level of communication, this group behavior manifests itself in an affective delimitation from outsiders, which is characterized by internalized antipathy and pre-emptive mistrust. The internal cohesion of such a social group is, hence, primarily ensured by its external relationships, especially towards other social groups. In this sense, the formation of a shared group ideology depends to a large extent on whether a social group achieves a communication-relevant differentiation between in- and out-group.

To represent the identification with a group, we introduce a new type of node, namely communities, as $c \in \mathcal{V}_c$. Users' being members of a community is expressed via edge type τ_c . The method by which these edges are derived inductively is in principle arbitrary. Nonetheless, some considerations have to be made. To get an idea of the social structure embedded in the communication graph \mathcal{G}_c , we explicate the connections that can be derived from it to build a social graph \mathcal{G}_u . Tracing paths along both τ_r and τ_s as a conjunction of edges allows us to query \mathcal{G}_c in order to identify connected users. We assume a directed edge from user u_1 to user u_2 to exist if u_1 has engaged in connective communication with respect to a message m authored by u_2 . Please note that edges between users are, thus, not observed directly (e.g. via explicit following relationships) but are indirectly derived from communication. This is indeed an important distinction since follow relations often say little about one's social identity. Several studies indicate that subscribing to people who belong to different ideological camps happens frequently [Bar*15; BMA15; Gue*18]. In contrast, we interpret onward dissemination of content as an expression of assumed relevance for one's own social group.

Based on \mathcal{G}_u , we perform community detection to derive the membership edges. We want to emphasize that we assume group identity behavior to only be derivable from acts of connective communication. Opposed to this, sending a message initially does not tell us anything about the social relations or intentions of an author. Therefore, we utilize a directed variant of Leiden's community detection algorithm [TWVE19] such that users are assumed to form a community if they react similarly to a given set of messages which is in line with the systems-theoretical view of communication, and not action, being the elementary unit of social systems.

After determining community membership, we transfer these new edges into the latent information space by training a projection $\mathcal{P}(\mathbf{u}, \tau_c)$. Unlike the original community detection, where membership is determined in a binary fashion, projection allows us to locate it on a spectrum of real-valued numbers. Informally speaking, the distance $\delta_{\mathcal{P}}(u, c \mid \tau_c)$ of projected user u to community c , in terms of the previously described semantics of the information space, depicts how strongly a user represents the shared characteristics of a community.

We now want to translate this relationship to our simulation procedure by taking the distance as a measure of how strongly a user identifies with a group. Following prototype theory, we, hence, describe membership of entities to certain categories by a quantitative gradation [Ros73]. Prototypes are central normative category elements to which other entities can only be determined relatively with a more or less large distance. Thus, prototype semantics assume that ideal categorizations never take place exactly, but can only gradually approximate an abstract prototype. Thereby, we calculate the strength of group influence $\eta \in [0, 1]$ by the relation between a user's prototypicality with respect to in- and out-group:

$$\eta = \frac{2\delta_{\mathcal{P}}(u, c_u \mid \tau_c)^{\kappa}}{\delta_{\mathcal{P}}(u, c_u \mid \tau_c)^{\kappa} + \delta_{\mathcal{P}}(u, c_m \mid \tau_c)^{\kappa}}, \quad (3.4)$$

where $\kappa \in \mathbb{N}$ is an ideological sharpness parameter and c_m is the representation vector of the community from which a message m originates.

We let this conceptualization of group identity influence communication by shifting the latitude of acceptance depending on whether a new input originates from the in- or an out-group. Formally, we define group influence as a weight on the latitude of acceptance:

$$\mathcal{L}_2(\tau_r(u, m)) = \frac{\eta\lambda^{\mu}}{\delta_{\mathcal{P}}(u, m \mid \tau_r)^{\mu} + \eta\lambda^{\mu}}, \quad (3.5)$$

The influence becomes stronger the more an individual identifies with their own group's views and the less they identify with an out-group from which the input originates. Note that this formula completely abstracts from the author of a message. Instead, messages are viewed as being representative of their specific social group. This is a conscious modeling decision to carry on with the idea that out-group members are met with preemptive mistrust. Finally, also consider the fact that $\eta = 1$ in case a new input message stems from a user's in-group. As a result, for in-group communication \mathcal{L}_2 is equivalent to \mathcal{L}_1 .

In the context of systems theory, the structural influence of group dynamics is crucial in establishing the distinction between self-reference and foreign reference within an ideological system. This relationship is realized in the proposed model by referring to system prototypes. Hence, the structural coupling relationship can be characterized by shared access to a collectively established structure, signified by the interdependent positioning within the information space. Consequently, this mode of referential world orientation of the system helps reduce complexity, as authors, particularly unfamiliar ones, are not identified as individuals but as members of their own or of an opposing social group.

This, however, does not elaborate on how the social system identifies a self, i.e., how it recognizes the connectivity of operations within the system and continuously reproduces the difference between system and environment. In the case of ideological systems, this process is achieved equivalently to functional systems, "through a binary code that assigns a positive and a negative value, excluding third possibilities" [LC00]. It is essential to note that this internal/external relationship of the code should not be confused with the difference between system and environment. Rather, they depict orthogonal concepts. In our case, the decision to utilize a piece of information is significantly influenced by the structural reference to system representations, and thereby by a differentiation between in- and out-group. However, the difference between system and environment describes the external boundary, while the binary coding of a positive and a negative value represents a mode of internal boundary maintenance.

The positive value thus denotes the connectivity of operations inherent in the system, i.e., what can be processed within the system. The negative value, on the other hand, reflects the conditions under which the positive value can be applied. The code is thus a two-sided form, the inside of which presupposes the existence of an outside. In the case of ideological systems on digital platforms, this code is specified as the distinction between information and non-information, expressed by \mathcal{L}_2 . Information is thus the positive value with which the system designates the possibilities of its own operation. But in order to work out the possibility of classifying something as informative, it must also be able to specify the counter-value in terms of irrelevance. Without such a reflexive value, the system would be at the mercy of the entirety of network communication and thus unable to distinguish itself from the environment. As a result, it would lack the basic operational capability of a system to reduce complexity and organize selections.

The core difference between information and non-information is employed to discern how potential connections are disrupted, enabling the formation of boundaries and the development of complexity within ideological sub-systems enclosed by self-imposed borders through a unique form of communication. A productive differentiation is required, which, under favorable conditions, leads to the emergence of the system. With the development of dissemination technologies being the critical achievement in the differentiation of digital mass media, ideological systems evolve specific planned modes of communication based on this technological foundation to support their autopoiesis. This operative closure allows the system to generate its operations autonomously, making it independent of maintaining interactional contacts with the environment and instead focusing on the system's intrinsic distinction between self-reference and foreign reference.

"For this [the code difference] uses a distinction - not a principle, not an objective, not a statement of essence, not a final formula, but a guiding difference which still leaves open the question as to how the system will describe its own identity; and leaves it open also inasmuch as there can be several views on the matter, without this 'contextuality' of self-description hindering the system in its operating." [LC00].

In our model, this potential of varying perspectives on the code difference is ensured by continuing to refer to the distance between user and message in \mathcal{L}_2 , meaning the distinction between information and non-information is ultimately determined individually.

Further note that cross-cutting communication can be seen as external stimuli or inputs that the system processes and adapts to its own internal logic. Messages from "outsiders" can either be

ignored, rejected, i.e., classified as non-information, or be integrated into the existing framework of beliefs and values, depending on how it aligns with the system's core principles. As a result, it is plausible for a message to be considered cross-cutting while establishing a system-internal context of meaning and disseminating it to one's circle of recipients as an element of system reproduction. From a systems theory perspective, it is inconsequential whether the message's dissemination signifies agreement or disagreement, as long as it is classified as informative within the system. We convey this property by allowing cross-border communication to be met with rejection, however not necessarily but only with increased probability. Instead of making a binary distinction between in-group and out-group, i.e., asserting the difference between system and environment, we place the distinction between informative and non-informative on a spectrum of distances.

This nuanced understanding lays the groundwork for recognizing the complexities associated with user affiliations in an online social network, particularly in border areas. In these regions, user affiliations are often unclear due to their extensive connections with multiple communities. Even more crucially, real-valued distances consider the distinctive relationships formed between diverse ideological systems within an online social network.

In the context of social polarization, the geometric arrangement of entities serves as an illuminating lens, casting new light on the unique complexities of ideological systems. Each system and its members inhabit a common information landscape, their positions determined by the inherent logic and dynamics of their ideologies. The semantics of this landscape are not arbitrary, but instead derived from the calculated distances between these systems. This spatial metaphor enables us to discern and interpret intricate relationships with a heightened clarity (cf. Figures 3.1 and 3.2). Some ideological systems lie closer together geometrically, signifying shared values, perspectives, or mutual influences. In contrast, others are positioned further apart, reflecting fundamental differences and potential points of conflict. Through this lens, the interplay of ideological systems transforms into an intelligible map of distances and proximities – a vivid topography of agreement and dissent. This perspective not only improves our understanding of these systems and their members but also provides a vital foundation for ongoing investigation into social polarization in the digital age.

Finally, note that while the distinction between information and non-information is also adopted in case of individual demarcation, \mathcal{L}_1 only addresses the question of how the overarching system achieves operational closure. Hence, the potential segregation of user groups in this case merely counts as a by-product of the relational process, for instance via epistemic gaps, with the respective communities not realizing communicative differentiation and closure due to a missing reflexive demarcation from the environment.

Summarized, in essence, ideological differentiation, and specifically the phenomenology of echo chambers, is not primarily captured by complete communicative isolation, but by the fact that ideological systems tend to realize specific forms of communication that make their own reproduction more likely. A resulting topological segregation of groups, however, is only a very probable but not a necessary product of the autopoietic process. In our view, therefore, classifying echo chambers solely according to the extent of internal vs. external communication misses the core of the problem of social polarization. Rather, the focus should be shifted towards developing ways of identifying system-internal structures that, in particular, make it possible to assess which cross-border communication is integrated into system reproduction. Both active

confrontation and deliberate dissociation from external content can contribute to system preservation, as can the consideration of such content for system-internal horizon broadening. Only in this way can an ideological system counteract the compulsion to constantly redefine its relationship to the environment by means of an adaptive adjustment of communication expectations due to the constant reconfiguration of the communication spectrum.

3.4.2 Recommender Systems

In online social networks, the transmission of information, i.e., the continuous linkage between ego and alter, are especially crucial. Unlike real-world interaction systems that can spontaneously and dynamically emerge from everyday situations and are spatially constrained, communication on social platforms is predominantly subject to technical limitations.¹² Therefore, social networks inherently belong to mass media, which Luhmann defines as the collection of institutions "which make use of copying technologies to disseminate communication" [LC00].

Essentially, distribution technology serves as the medium through which communicative forms are established, acting as the foundation for specific operations. Characteristically, such distinct forms of communication enable the system – in this case, the overarching system of a particular platform – to achieve operational closure. However, digital communication possesses unique characteristics that are challenging to align with Luhmann's mass media concept from the 1990s, developed long before the internet gained its current prominence. Technological advancements have since enabled the emergence of new media types, employing vastly different reproduction methods than traditional distribution media like newspapers or television.

The fact that anyone can sign up for a platform makes this *audience turn* appear as an egalitarian means of participating in expansive communication, empowering former information consumers to actively contribute to public opinion formation. This leads to connected communication occurring within the social system itself, as opposed to other mass media that may only incite the creation of independent systems in response to perceived communication. For instance, a married couple discussing news they just watched on television exemplifies the latter. In the case of social networking sites, however, the wife may choose to respond to a message herself, actively engaging in the communication process. Hence, the dissolution of the producer-consumer dichotomy is key to comprehending digital network communication dynamics.

Consequently, a significant outcome of the increasingly platform-dominated media landscape is the substitution of traditional gatekeepers with new distributors in the form of automated recommendation algorithms. The occurrence of communication is no longer solely under the social system's control but relies heavily on the associated technological channel. This implies that potential recipients are not accountable for coordinating their consumption interests, like purchasing a specific newspaper at a newsstand; instead, recommender systems' personalization logic claims to manage this for them.

Paradoxically, the enhanced possibility of participation is accompanied by passivity in consumption. In contemporary times, internet users frequently do not actively search for news; instead, personalized feeds automatically select news for them without further manual intervention. As

¹²The exact implementation of the technical basis and a deeper discussion of algorithmic approaches shall not play a role here. Instead, we want to place it conceptually within the theoretical framework presented here. A detailed description can be found in [DZ21; DZ23]. See also Chapter 2 for some insights.

a result, individual horizons of information are largely subject to algorithmic control. Recommender systems can expand, condense, or entirely close off communication channels, thus controlling or promoting the dissemination of specific information. Consequently, authors are no longer responsible for making assumptions about appropriateness and acceptance, even though these factors continue to be crucial in differentiating particular sub-systems.

In digital mass communication, the standardization degree is considerably lower than in traditional distribution media programs, leading to a significant differentiation in information distribution. This enables the control of an initial surplus of communication possibilities so that specific forms of social system self-organization are enforced. Problematically, it is usually impossible to determine from an external perspective which parameter combination was ultimately responsible for delivering certain content, as the underlying decision logic remains a black box.

3.4.2.1 Self-reflexivity of Recommender Systems

It is crucial to emphasize that the technologically mediated content dissemination is not an operation within the social system itself: "Not everything which is a condition of possibility of systems operations can be a part of the operational sequences of the system itself." [LC00]. The platform's reality and the ideological systems stemming from it can only be perceived through the communication occurring within it. However, the communication medium so fundamentally influences the addressees of a message and the potential for connective operations that the importance of technological influence on the social phenomena examined here cannot be dismissed.

As extensively discussed in this thesis, successful content distribution, and thereby the facilitation of digital communication, heavily relies on tailoring information to specific audiences, as only then can a social network effectively reproduce itself as a system. The vast amount of information available is no longer filtered by humans (e.g., editors, producers, or program directors), but rather by technology. While the social system persistently applies new communication to previous communication results, the technological apparatus must ensure that connectivity conditions are adequately present. The algorithmic delivery, hence, must continuously generate and process irritations, converting information into non-information.

To accomplish this, a personalized recommender system necessitates an internal representation of the communication process. In other words, it must be structurally oriented towards constructing the reality of the various social systems that reproduce themselves through the technological medium. As a consequence, one primary distinction between traditional mass media and digital social networks¹³ lies in the latter's reliance on approximating access to real-world latent structures established within a social system.

As a result, online social networks not only function to offer a form of public memory from which new communication can be generated through structural reference – which, according to [LC00], is the primary social function of mass media. Instead, they rely on maintaining a representation of this memory themselves to effectively function. Thus, technological apparatuses of

¹³While it is valid to consider each media outlet, be it a traditional newspaper, a TV channel, or a digital platform, as an operationally closed system from a systems-theoretical perspective, it is important to recognize the structural coupling between traditional media and digital platforms. As these outlets share content and influence each other, their interactions help shape the overall media landscape. Thus, the analysis of the impact of recommender systems and digital platforms on communication processes should ultimately take into account the specific web of relationships between different media systems.

digital platforms are not merely media because they convey information from those who know to those who do not. Instead, they are media insofar as they supply and utilize background knowledge and update it in a specific manner through their technological mode of operation. Unlike traditional mass media, platforms do not partition their distribution spectrum into pre-defined programs based on assumptions about the target audience. Rather, the algorithms' flexibility and adaptability ensure that the medium itself assumes a steering function primarily determined by its capacity for extensive differentiation.

In summary, a recommender system not only reconstructs social structures to deliver personalized content but also actively influences them by substantially intervening in the relational process of users. As a result, it becomes a complex adaptive system that ensures its own continuity by securing its subsequent operations through a reflexive reference to its internal state. Its reproduction is achieved by facilitating ongoing communication in the social network in a specific manner. However, the mode in which this occurs may be led by design principles that deviate from individual user interests or societal considerations. Indeed, the recommender system does not rely on building an accurate representation of real-world structures. Instead, in a constructivist manner, it creates its own approximation of external reality, which serves exclusively for its own reproduction. It is important to note that certain parameters of the internal functional logic still depend on couplings with other systems, particularly the economic system. Therefore, the recommender system can act as a lens that deliberately distorts actual social structures. Its disruptive potential lies in that this distortion can increasingly become the truth, as communication in social systems is directed accordingly, and structural adaptation takes place. The crucial question now is to what extent this derangement actually manifests itself in social settings.

3.4.2.2 The Role of Recommender Systems

Although recommender systems may not operate based on exact inferences, their data-based assumptions about content suitability generally provide certain benefits. Their imprecision, in fact, offers the advantage of not creating overly restrictive loops, preventing communication from being immediately blocked by failure and contradiction, but rather allowing it to seek a receptive audience and explore various possibilities. By avoiding excessive prescription, recommender systems may enable a diverse range of content to circulate within social systems, potentially fostering dynamic communication environments where individuals can encounter different perspectives and experiment with new ideas. They have the potential to serve as crucial connectors that shape and structure the communication process, allowing for diversity and adaptability amid a constantly evolving information landscape.

Online social networks thus provide a vast space for dissemination; but it is not unlimited, as not all information can be distributed with equal success to different recipients at specific times. Recommender systems must continually adapt and modify the structural conditions of couplings between individuals, altering the schema offered according to their internal logic. Thereby, they must ensure the ongoing transformation of information into non-information while taking into account the receptivity of specific audiences. As black boxes, however, it becomes difficult to determine the extent to which communication channels have been technologically constrained or expanded. This, in turn, makes understanding the transformation of information into non-information challenging. The primary issue is not digital corporations withholding the algorithmic foundation of their dissemination technology, but the anonymization of social redundancy

which renders the control effect of recommendations opaque. With social redundancy we mean the repeated interactions among the same individuals in various roles, such as authors, recipients, or disseminators of messages, resulting in an excess of shared ideological fragments in communicative circles.

Specifically, message creators may not know who received their messages, while receivers might be unaware of other – possibly contrasting – messages circulating beyond their information horizon. As a result, the generation of such technological constraints represents the core problem of filter bubbles. Recommender systems thus adopt a distinct role in the context of ideological systems in online communication by being accountable for creating and maintaining social redundancy, which is particularly vital for stabilizing intra-group communication.

Expanding on this concept, the temporal nature of information is highlighted. Luhmann suggests, "Information cannot be repeated; as soon as it becomes an event, it becomes non-information. A news item run twice might still have its meaning, but it loses its information value" [LC00]. Yet, within ideological systems, the dissemination of known content can still hold value. The association between the message and its disseminator can act as a signal for alter. Even if the message's content is familiar, its sharing by a particular individual or group remains informative. In social networks, repetition can add layers of meaning beyond the initial content. These repetitions, acting as social cues, strengthen group cohesion and amplify the connectivity of the ideological system. This transformation of new information into reinforcing redundancy characterizes the dynamics of *echo* chambers.

In summary, we argue that algorithmic dissemination may lead to a progressive condensation of information channels, which in turn can undermine undistorted information gathering. One of this work's central thesis is that the necessary restriction of overwhelmingly large amounts of information is executed by the personalization logic of recommender systems, particularly along the operational boundaries of different ideological systems. In our view, as the derived numerical representations of users and content may fail to fully encapsulate the subtleties of individual needs and intentions, personalized news feeds primarily reproduce ideological contexts of meaning through a (deliberately) inaccurate duplication of structure formation and reference. Although ideologically congruent systems form in the context of other communication channels without recommendation technology's influence, content personalization potentially condenses, accelerates, and amplifies the process of ideological difference formation. Again, however, it must be emphasized that this does not imply a complete communicative decoupling of ideological systems. Rather, the deliberate dissemination of contrastive content can also increase the likelihood of reproduction if it is met, for example, in the form of a collective counter-reaction.

While dissemination technologies do indeed operate with respect to the motivations of communicating individuals, even making them the starting point of their internal logic, they, however, do not necessarily understand motives as psychological or physiological causal factors, but merely as a means of accounting for individuals in the communication process. From an operational standpoint, the reasons for action, which are explicitly linked back to individuals through personalization, are artifacts of social communication. Platforms are therefore primarily interested in how individuals can participate in generating further communication, regardless of their thoughts or the consequences for them or society.

The fact that ideological demarcation and the resulting polarization patterns can have negative social consequences initially plays only a subordinate role for the platform operators. It is essential to acknowledge the potential trade-off between the economic interests of platform companies and the liberal democratic aspiration to guarantee and ensure access to a balanced diet of information. The news feed, as the central interface between people and platforms, relies on predictive algorithms that extract numerous parameters from users' behavioral surpluses to calculate the personal relevance of thousands of eligible messages [Zub19]. As should have become clear by now, these algorithms typically favor messages from people with whom users have interacted in the past, messages that have attracted significant interest from others, and messages similar to those that have been of interest to the user previously.

In this context, reality on social platforms is structurally coupled with the economic system through the channel of dissemination technology, as a company's economic interests significantly influence the way messages are disseminated. Digital platforms aim to create a closed loop that feeds on the needs and inclinations of its users, amplifying and then potentiating them. In this context, *needs* primarily refer to expected values extracted from behavioral data, which may not necessarily align with a person's actual needs. As a result, personalized content might seem to fulfill these needs and increase the user's (emotional) engagement in the short term; however, it remains contested whether positive effects persist in the long run.

In this sense, digital mass media seemingly offer a freedom that they do not necessarily live up to. True freedom depends on the cognitive conditions of observing and describing alternatives with open, decidable, but unknown futures. While psychic and social systems may generally empower themselves to choose, the deliberate restriction of the distribution spectrum undermines the freedom-constituting potential that digital communication could actually realize. Instead, platforms can produce *value change* through privilege and enclosure. Minority opinions, for example, can be prominently displayed because they are spectacular, conflictual, or deviant, potentially triggering the *spiral of silence* identified by [NN93].

The individual tailoring of information horizons creates the impression of a consensual world and the creation of social redundancy eliminates the need for individuals to directly distinguish their view of reality from that of their environment. The social support of like-minded people acts as a motive to justify one's own view, allowing it to be regarded as universally valid, or reality par excellence. These communicative conditions can foster the development of fundamentalisms of all kinds. As people assert, "This is my world; this is what we think is right", the resistance they encounter can be a motive to escalate, potentially leading to radicalization without necessarily causing doubt about reality.

3.5 Simulation Framework & Empirical Validation

Having thoroughly examined the theoretical underpinnings of social polarization, the dynamics of individual and ideological demarcation, and the role of recommender systems, it becomes imperative to transition from theoretical exploration to empirical validation. The theoretical discourse thus far has underscored the complex interplay of social and technological factors in shaping social polarization. However, the mutable nature of group boundaries and membership within ideological social systems, the path-dependent elements, and the dynamics of social influence necessitate a more rigorous and dynamic methodology for studying this phenomenon.

In this section, we hence introduce a simulation-based approach, designed not only to validate these theoretical findings but also to provide a more nuanced understanding of the dynamics at play. This approach, which capitalizes on the multidimensionality of latent information spaces, is capable of capturing the subtleties and interdependencies of opinion formation and spread over time. It also integrates the influence of recommender systems, thereby enhancing the model's realism and expanding its applicability.

Through this simulation method, the aim is to address the limitations of previous methodologies and provide a more comprehensive understanding of social polarization, one that captures not only the process of demarcation per se but also the construction of complexity in the representation of the environment on the level of social systems. The following section will first position the proposed model in the research landscape of social polarization. Afterwards, we will detail the specifics of our simulation procedure and how we validate it.

3.5.1 Literature Review & Model Comparison

The complexity of social polarization, particularly its manifestation in online social networks, warrants a multi-disciplinary approach that incorporates insights from fields such as sociology, psychology, and computer science. To this end, an array of methodologies has been utilized, including experimental studies, social network analysis, mathematical modeling, game theory, and simulation-based models.

Experimental studies offer a key starting point, providing controlled environments for observing individual and group behaviors [FH09a]. Although they may lack the scale and complexity of real-world social interactions, these laboratory-based studies afford invaluable insights into the micro-level dynamics of social polarization. They establish an empirical foundation that allows for evidence-based development and refinement of theoretical models. Complementing these empirical investigations, Social Network Analysis (SNA) offers a way to map and measure relationships and flows between people, groups, and other connected entities [Wat04; Bor*09]. SNA visualizes and quantifies the propagation of information (or misinformation) across a network, thereby contributing to our understanding of how ideas spread and polarization intensifies.

Mathematical modeling approaches, such as the Deffuant model [Def*00] and the Hegselmann-Krause model [HK02], provide quantitative and predictive capabilities for exploring opinion dynamics and social polarization. These models use mathematical formulas to represent the social interactions that lead to polarization, offering a theoretical perspective on the mechanisms at play. Furthermore, game theory provides a strategic framework for modeling interactions within social systems [Son*20; MR21]. Game theory captures the strategic decisions of individuals who adjust their opinions and behaviors based on perceived benefits and costs, helping to elucidate the individual-level decision-making processes that contribute to larger polarization phenomena.

Building upon these approaches, simulation-based models have been developed to represent the dynamic nature of social interactions. Agent-Based Modeling (ABM) is one such approach, wherein each individual or *agent* is modeled as an independent entity with its own set of rules and behaviors [RG19]. ABM simulates how individual behaviors collectively contribute to macro-scale societal phenomena like social polarization. Further, complex contagion models represent a more advanced form of simulation that captures the influence of broader ideological and cultural patterns on individual behavior [Tö18; VLP19]. These models accommodate the notion

that individuals are influenced not only by their immediate connections but also by wider societal forces, representing an attempt to capture the complex, multi-layered nature of real-world social networks.

Despite these advances, all these methodologies, including both ABM and complex contagion models, implicitly or explicitly assume that macroscopic polarization effects primarily result from interactions between individuals, either locally or influenced by wider societal patterns. This research argues that this focus on individual interactions may not sufficiently account for the formation and operation of social groups as autonomous entities that differentiate themselves from other groups. In response to this perceived gap, we propose a novel simulation approach to modeling social polarization, wherein social groups are conceived as distinct types of social systems that establish a difference between themselves and their environment. Members of these social systems employ shared means of communication primarily to differentiate in-group members from members of potentially multiple out-groups.

This approach offers several potential advantages over the existing simulation methodologies. It addresses the role of shared cultural and communication patterns in driving polarization, accommodating group effects and wide-ranging social influences. By treating the social group as a unit of analysis, it simplifies the modeling process by avoiding the need for complex modeling of multiple types of connections. Moreover, it emphasizes the mechanisms of in-group/out-group dynamics and group identity formation, key drivers of social polarization according to sociological theories.

Our model, utilizing latent information spaces derived from a communication graph, additionally allows us to incorporate the distribution patterns of messages into our analysis, providing a richer and more detailed view of social interactions. Unlike traditional one-dimensional attitude spaces, latent information spaces can represent entities in a multi-dimensional space, capturing a wide range of ideologies and attitudes, and modeling scenarios where more than two opposing groups are present.

As such, the multidimensionality of latent information spaces allows for the depiction of more complex relationships and semantics. For instance, it can help capture the nuances of ideological shifts, the spread of ideological groups, and the dissolution of existing ones. Our model also integrates the influence of recommender systems, offering deeper insights into the technological influences that drive social polarization. This integration enhances the model's realism and expands its applicability, allowing us to model and analyze a wide range of scenarios, from highly ambiguous settings to the formation of echo chambers.

Another advantage of latent information spaces is their ability to track changes over time with respect to multiple dimensions. Observing how entities move within these spaces, we can gain insights about complex migration patterns, revealing the subtleties and interdependencies of opinion formation and spread which, in turn, may inform the design of recommender systems and other interventions. Furthermore, our model eliminates the reliance on additional contextual data, simplifying the modeling process and increasing its flexibility and scalability.

In conclusion, the use of latent information spaces offers a more nuanced approach to studying social polarization. By capturing complex dynamics, facilitating the study of more than two opposing groups, and tracking changes over time, our model provides a powerful tool for understanding and addressing the challenges posed by social polarization. The interpretation of these

spaces is a critical aspect of our model, providing a multi-faceted understanding of the dynamics of social polarization. Each position in the latent space represents a unique combination of factors influencing an entity’s behavior. Observing changes in these positions over time, we can study the dynamics of social polarization, such as social reinforcement or polarization.

3.5.2 Simulation Procedure & Validation

Our primary objective is to investigate under what conditions do ideological systems foster enough connectivity to efficaciously self-replicate. To this end, we take into account the technological variables of communication, which play a pivotal role in the framework conditions for deliberate communication. Our goal is to scrutinize how the condensation of initial communication possibilities, brought about by algorithmic message distribution, increases the probability of subsequent connectivity within the larger social system and the associated ideological sub-systems.

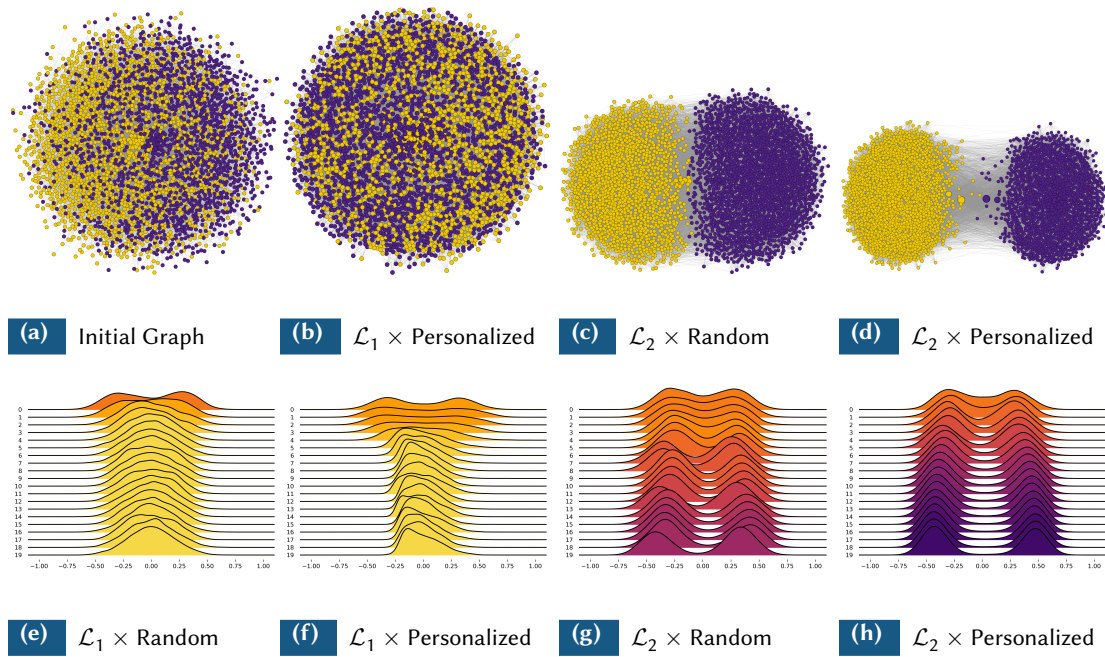
We initiate our approach by employing either a synthetically produced or an empirically observed real-world communication graph. We hypothesize that this graph hosts specific communication patterns enabling us to approximate its structures. Following this, we model the progression of network communication over multiple epochs in the simulation, a process that is bifurcated into two primary steps: generating message recommendations for each user and, subsequently, modeling connective communication predicated on those recommendations.

While the choice of the recommendation algorithm is, theoretically, arbitrary, we have for this study, derived its logic from the latent semantics of the machine learning model presented. Our approach, incorporating established concepts from the domain of recommender systems research, is detailed in [DZ23]. The principal aim of this approach is to construct a system that mirrors real-world social networks closely and scrutinizes the influence of content personalization on social polarization. To achieve this, we explore both personalized and non-personalized message dissemination and probe how these methodologies impact the relationships between entities over time. This line of investigation allows us to discern whether communication channels tend towards progressive densification, as per the filter bubble hypothesis, and to evaluate the degree to which targeted user engagement is requisite for the autopoiesis of ideological systems.

In relation to human communication, we have established a crucial distinction between self-referential and group-referential behavior (\mathcal{L}_1 vs. \mathcal{L}_2). Self-referential connective communication is propelled by individual cognitive needs, whereas group reference emerges from a collective identity that unites members of a social group. This differentiation is paramount for grasping the diverse roles individuals can assume in the formation and maintenance of social systems, particularly concerning the differentiation of sub-systems.

As the simulation traces the temporal evolution of communicative processes, we presume that the relationships between entities are in a constant state of reconfiguration, resulting in a dynamic communication graph. Specifically, the generation of connective communication, i.e., when a user accepts a system-recommended message, leads to the addition of new edges to the communication graph¹⁴. This process can forge new connections between previously unconnected users, thereby influencing both the user graph and community detection respectively. We hence

¹⁴We also allow for the removal of existing edges based on specific heuristics [cf. DZ23].

**Figure 3.1**

Exemplary visualizations of graph topologies and latent space polarization for the partisan scenario. The graphs shown for each condition (figs. 3.1b to 3.1d) are taken from one randomly selected simulation run at final epoch 100. While in the initial graph (fig. 3.1a), the two communities are strongly intertwined, all conditions are able to clearly separate them, with $\mathcal{L}_2 \times$ Personalized achieving the strongest delineation. Condition $\mathcal{L}_1 \times$ Random is omitted because the graphs shows no notable effects. Below (figs. 3.1e to 3.1h), polarization in latent space is depicted in the form of joyplots (also known as ridgeline plots) giving a sense of the evolution of polarization over time (y-axis). The plots show the kernel density estimations of a one-dimensional principal component analysis (PCA) on the latent representations of user nodes. The color gradient from yellow (low) to violet (high) depicts the PCA's explained variance. The results reveal that the slight separation at the start cannot be maintained by conditions $\mathcal{L}_1 \times$ Random and $\mathcal{L}_1 \times$ Personalized, while the remainder increase bipolarization, with $\mathcal{L}_2 \times$ Personalized showing the most pronounced effects.

view ideological groups as fluid structures with evolving boundaries and memberships over time. The updated constellation of links informs subsequent simulation epochs, potentially leading to the adaptation of both user behavior and recommendation generation.

Summarized, to validate the proposed framework, our primary goal is to demonstrate that the choice of human decision function significantly influences whether a social group stabilizes over time or disintegrates. The selection of the decision function hinges on the specific scenario, which could be either exclusively self-referential, group-referential, or a blend of both. Furthermore, our model allows for the variability in the number of identified communities, thereby capturing not just two-sided polarization effects but also the dynamics of multiple competing social groups. In addition, we aim to clarify the relationship between human communication and the choice of recommendation algorithm. Specifically, we seek to examine whether the personalization logic underlying social media recommendations can contribute to group stability, and potentially

exacerbate centrifugal tendencies.

The main independent variables are, first, the human decision function $\mathcal{L} \in \{\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_{1|2}\}$ and, second, the choice of recommendation algorithm $\mathcal{A} \in \{\text{personalized}, \text{random}\}$. Note that $\mathcal{L}_{1|2}$ depicts some combined application of \mathcal{L}_1 and \mathcal{L}_2 . By *random*, we refer to message recommendations that are not calculated with respect to any personalization logic, but are rather selected arbitrarily from any graph user. This creates a two-dimensional $\mathcal{L} \times \mathcal{A}$ study setup.

We apply this setup to synthetic and real-world communication graphs with varying degrees of group integration and polarization. For a detailed explanation of our method for creating synthetic graphs, and further analysis regarding the variation in community-internal and external connectivity, as well as the relative community sizes, see [DZ23].

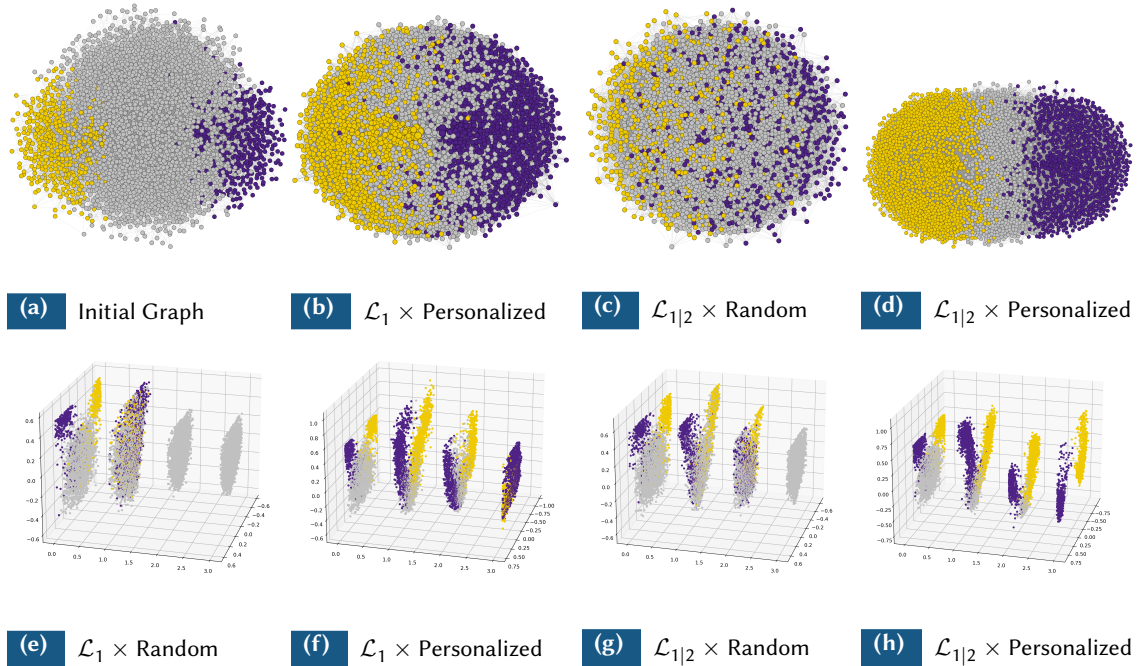
3.5.2.1 Synthetic Scenarios

To create a controlled environment that validates our approach, we first turn our attention to synthetic communication graphs (see [DZ23] for a detailed explanation of the employed method to create synthetic graphs). This method offers the researcher complete control over parameter selection and the initial setup of connectivity, thereby shaping the initial state of communities within the network. It provides a flexible platform to probe our framework from different angles and perspectives, ensuring that our simulation accurately reflects the theoretical constructs we aim to model. This strategy enables us to manipulate the variables of interest while keeping other potential confounding factors constant, thus establishing a more robust cause-effect relationship.

Our primary research question explores the relationship between ideological delineation and algorithmic dissemination, and to this end, we focus on two distinct scenarios. The first is a partisan scenario where participants initially align themselves with one of two ideological camps, although shifts in this alignment are not precluded as the simulation progresses. The second scenario encapsulates an undecided case, with two diverging viewpoints emerging on the periphery of the discourse, while the majority of participants have not yet cemented their position. Initial graph topologies for each scenario can be found in Figures 3.1a and 3.2a. For a more detailed discussion on these findings, the reader is referred to the respective paper [DZ23]. This section provides a high-level summary, sufficient for the scope of this thesis, of the simulation results in the context of our proposed framework.

Our results underscore the influence of the initial debate structure and individuals' ideological affiliations on the incipience of polarizing effects. When the social graph is pre-structured along ideological lines, the polarization driver \mathcal{L}_2 in our simulation leads to debate fragmentation, even in the presence of a broad recommendation space ($\mathcal{L}_2 \times \text{Random}$). This finding mirrors real-world debates that often evolve along pre-existing ideological fault lines such as party affiliations.

On the other hand, even though personalized message distribution densifies certain channels, the initial separation appears to be transient when individual demarcation alone is considered ($\mathcal{L}_1 \times \text{Personalized}$). This suggests that technological factors alone may not account for the bifurcation of discussion spaces. In cases where homophilic behavior intersects with personalization technologies, the outcome is not a marked creation of epistemic gaps as predicted by the echo chamber hypothesis. Instead, it leads to a postponement in consensus building. Our findings

**Figure 3.2**

Exemplary visualizations of graph topologies and latent space polarization for the divisive scenario. The graphs shown (figs. 3.2b to 3.2d) are each taken from one simulation run at epoch 50. Initially (fig. 3.2a), two small communities (depicted in violet and yellow) are integrated to some degree into a larger community (depicted in grey) while forming at opposite edges of the social graph. In condition $\mathcal{L}_1 \times$ Personalized, both peripheral communities spread out, while in $\mathcal{L}_2 \times$ Random, they become further integrated into the grey community, thus loosening internal connectivity. Concerning condition $\mathcal{L}_2 \times$ Personalized, we observe pronounced centrifugal tendencies with condensed structures forming separately in the focal points of the respective communities. Below (figs. 3.2e to 3.2h), polarization in latent space is depicted. Due to the presence of not only two, but three communities, we depict latent user distributions in form of a three-dimensional scatter plot. Concretely, we show the migration of user distributions by plotting a two-dimensional PCA over time at epochs 1, 25, 50, and 100. Clearly, only condition $\mathcal{L}_2 \times$ Personalized shows bipolarization while the remaining converge to consensus.

corroborate studies arguing that digital platforms can potentially bridge preference gaps over time, despite users' individual content selection [Bar*15; BMA15; Gue*18].

Comprehending social polarization more holistically consequently necessitates considering not only the technological apparatus but also the social demarcation aspect. However, the ideological leanings of communication participants may not always be as transparent as in the first scenario. Given the opacity of personal filter bubbles and individual decision criteria for communicative behavior, determining the reasons for the greater divergence in some debates remains challenging. The uncertainty surrounding the roots of these divergences – whether they originate from human communication, dissemination technology, or a blend of both – highlights the complexity in identifying the causal mechanisms of social polarization.

In our second scenario, where ideological sorting within a particular network structure is not readily apparent, the dynamics become more complex. Such topological preconditions are particularly to be expected in case of novel or multifaceted subjects, where initial discussions may be wide-ranging and the formation of personal opinions gradual. Our model indicates that even tightly knit communities on the extreme ends struggle to preserve internal stability under a broad recommendation spectrum and eventually become marginalized and assimilated ($\mathcal{L}_2 \times \text{Random}$).

Notably, the personalization of recommendations leads to the preservation of the two peripheral communities rather than the original dominant one. This outcome emerges even in the $\mathcal{L}_1 \times \text{Personalized}$ condition, where participants do not display overt group-referential behavior. We ascribe this phenomenon to the robust initial networking within marginalized communities, fostering sustained internal stability. However, the resulting consensus is pluralistic rather than fragmented. In case of $\mathcal{L}_{1|2} \times \text{Personalized}$, instead, the initially undecided community, characterized by a lack of firm ideological commitment, may be more susceptible to the sway of influential gatekeepers and personalized recommendations. As community members become more ideologically cohesive, they may adopt the \mathcal{L}_2 decision function, thereby contributing to the fragmentation of the debate space.

A key insight from our model is that social polarization only transpires when ideological delineation is complemented by communication and personalized recommendations. We deduce that the gradual demarcation of ideological groups stems from a complex interplay of social and technological factors. Therefore, to comprehend social polarization, we must scrutinize these facets conjointly. These outcomes correspond with the theoretical framework of ideological social systems, which posits that individual interactions are propelled by their ideological beliefs, preferences, and affiliations. The polarization patterns observed in our simulations can be interpreted as the byproduct of interactions among users with differing ideologies, coupled with the amplifying effects of personalized recommendation algorithms. This underlines the instrumental role that ideological social systems play in sculpting the dynamics of network communication and polarization.

Our simulations also underscore the mutable nature of group boundaries and membership within ideological social systems. As users encounter diverse messages and recommendations, their ideological affiliations may pivot, engendering shifts in the community structure and the degree of polarization. This fluidity accentuates the intricate interplay between individual ideological

beliefs, their selection of decision functions, and the impact of recommendation algorithms on the evolution of the ideological social system.

3.5.2.2 Real-world Scenario

As our exploration navigates from synthetic to real-world networks, we present an initial validation of our model using actual data. A communication graph harvested via the Twitter API, centering around the theme "Ukraine" spanning from November 1, 2021, to February 23, 2022, serves as the foundation of this investigation. Our model's sensitivity to data sparsity compelled us to amplify the graph's density. The revamped communication graph, comprising 4,946 users, 2,225 messages, and 75,970 edges, is designated as our target graph. With a pioneering subgraph, formed from the foremost 20% of edges in the target graph, we aspire to emulate the community structure of this graph as closely as feasible.

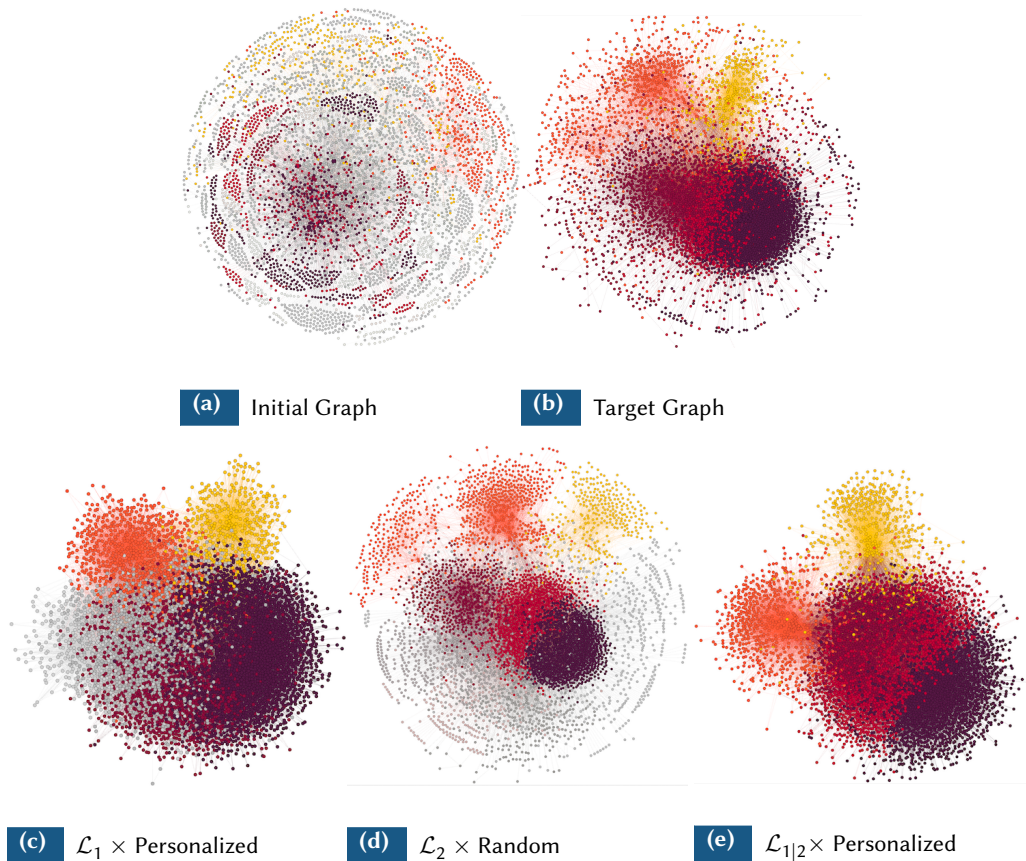
We employed the human decision functions \mathcal{L}_1 and \mathcal{L}_2 , introduced earlier. Again, we apply the hybrid condition $\mathcal{L}_{1|2}$, allowing us to investigate a scenario where the network exhibits a blend of both homophilic only and ideological behaviors. The recommendation conditions employed, personalized and random, were previously defined. In this real-world validation, we incorporated 'weak-tie' recommendations to our considerations. This new component enables us to more accurately mimic the dynamics and connectivity patterns that typify social media platforms.

The main objective of applying our model to real-world data was to assess its portability to empirical datasets. The empirical graph, in contrast to synthetic ones, unmasked more distinct subnetwork structures. The efficacy of our model was measured using homophily and modularity as key metrics. The results suggest that a confluence of personalized recommendations and group-referential decision behavior fosters the creation and preservation of homophilic structures.

Upon detailed temporal dissection of the network state, a discrepancy between \mathcal{L}_1 and \mathcal{L}_2 comes to light. The \mathcal{L}_1 function shows a propensity to quickly dismantle community structures, underestimating polarization tendencies while simultaneously only delaying consensus through personalized recommendations. Conversely, \mathcal{L}_2 function overstates the fragmentation of the information space. The target graph is hence better characterized by consolidated polarization rather than fragmented polarization.

To approximate this consolidation process, we introduced the condition $\mathcal{L}_{1|2}$, a combination of \mathcal{L}_1 and \mathcal{L}_2 . This function posits that not every community necessarily exhibits systemic behavior. Initial separations could be attributed to unsystematic epistemic differentials. Over time, smaller factions may be subsumed by larger, topologically adjacent communities, governed by the principle of hegemonic dominance. Moreover, it is conceivable that many actors may not harbor a fixed, ideologically-driven opinion initially.

This approach, we concede, does involve a significant level of abstraction and simplification. An accurate determination of the number of communities demonstrating systemic behavior would necessitate more comprehensive criteria, such as a thorough content analysis, observations of actors across diverse contexts, and longitudinal studies. Nevertheless, certain heuristics can provide an initially adequate approximation, for example, using political parties present in parliament for politically charged topics.

**Figure 3.3**

Visualizations of graph topologies on a real-world dataset. The initial graph (fig. 3.3a) consists of 20% of the earliest edges contained in the target graph (fig. 3.3b). While the initial graph has very loose connections and thus a large number of communities is identified, the target graph is much more condensed and contains only 5 communities. We calculated Jaccard similarity to the target network's communities to assign colors to the communities in the initial as well as the scenario graphs. All three depicted scenarios (figs. 3.3c to 3.3e) capture to some degree the properties of the target graph. For instance, they all identify the most clearly separated communities (in orange and yellow). However, at closer inspection $\mathcal{L}_1 \times$ Personalized shows too much and $\mathcal{L}_2 \times$ Random too little condensation. Although there are still obvious topological discrepancies to the target graph, we view $\mathcal{L}_{1/2} \times$ Personalized as achieving the closest approximation.

In terms of the quality of approximation to the target graph, the condition $\mathcal{L}_{1|2}$ offered the most precise predictions in the present study. This function ensures that the progression of separation remains within the bounds of the start and target graphs, avoiding the pitfalls of \mathcal{L}_1 and \mathcal{L}_2 functions, which swiftly overshoot or undershoot these bounds or cling too closely to the origin graph.

This study’s findings form a robust foundation for future work, pointing towards the application of hierarchical community arrangement for a more accurate representation of sub-community structures. The promising results affirm our model’s potential in encapsulating the intricate dynamics of network communication on social media platforms, paving the way for more nuanced and detailed future research.

3.5.2.3 Varying Recommendation Algorithms

Building on our application of the approach to real-world data, we now delve into the exploration of varying recommendation techniques [DZ21]. Understanding the way in which technology shapes our social connections through personalized information streams is fundamental. For instance, on platforms such as Twitter, the set of candidate users for recommendations is determined through structural links in the network graph. Subsequently, recommendations are generated from the set of postings that these users have interacted with in the past. In this subsection, we consider six such techniques, namely: Ego-network, Random, Recommendation, Boundary, Boundary-projection, and Community-intersection.

Ego-network This approach simulates one central aspect of the distribution patterns prevalent on social networks like Twitter as well as the basic diffusion principle of many opinion dynamics models. In these networks, individuals are linked through one-directional follow relationships. For any user, the algorithm recommends content by referencing these directly connected users and the postings they interact with.

Random To confirm that our proposed techniques do exploit the structural links between entities and do not merely benefit from increased diversity, we use a random selection strategy. This serves as a control condition, providing a benchmark against which we can measure the effectiveness of the other recommendation techniques.

Personalized recommendation This condition is similar to the *ego-network* approach in that it considers local neighborhoods to identify candidate users. However, instead of relying on graph edges, we inspect the neighborhood in latent space. For this, we apply $\mathcal{P}(\mathbf{u}, \tau_u)$ as a user recommendation process.

Boundary Recognizing the ideological structure of echo chambers, we argue that mitigating attempts should come from within these echo chambers. Previous research suggests that attitude changes are more likely to be incited by peers expressing unexpected stances. With this in mind, we hypothesize that statements challenging the ingrained beliefs of an echo chamber are often found in the boundary regions between communities, where external content seeps into insider discussions. For a target user u and their respective community c_u , we define candidate users as the boundary entities v_b that satisfy an edge relation $\tau_b(c_u, v_b)$.

Boundary-projection While connecting users with boundary representatives of their community may increase cross-cutting interaction, it also risks triggering backfire-effects. Exposure to counter-attitudinal information can prompt cognitive processes aimed at managing cognitive dissonance. We aim to mitigate these effects by selecting only boundary users likely to have a positive impact. This involves combining the user *recommendation* and *boundary* user selection ideas. From this, we learn $\mathcal{P}(\mathbf{u}, \tau_b)$ and extend the original recommendation generation $\mathcal{P}(\mathbf{u}, \tau_r)$ such that it follows the path over boundary users. This approach is built on the premise that this translation operation identifies boundary users who are best suited to accurately predict a user's interaction behavior.

Community-intersection The last technique we explore is the intersection across disparate communities, an approach that endeavors to create bridges between polar opposite groups. The underpinning principle here is to foster a more balanced conversation environment by deliberately facilitating connections between these disparate groups. This method employs an intersection operation to pinpoint users within a divergent community with whom the target user is likely to engage. It is worth noting that in some cases, our training dataset may not yield valid edges that satisfy a specific query. Despite this, the strength of our approach lies in its capability to approximate potential connections not readily observable in the existing data. To illustrate this operation, consider a target user u and their community c_u . We identify the most distant community by calculating the distance metric $\delta(\mathcal{P}(c_u, \tau_c), \mathcal{P}(c_v, \tau_c))$. Following this, we compute the intersection between the target user and the identified distant community. This process uncovers potential connections, fostering interactions that may otherwise remain unexplored, thereby potentially reducing polarization within the broader discussion space.

Upon introducing the various recommendation conditions, the ensuing discussion enables us to delve into the analysis of the outcomes. The results, as expected, shed light on the significant influence of technological filters in the diffusion of information within social networks and the reciprocal relationship between individual human integration procedures and diversification attempts.

The tendency for scientific research on opinion dynamics simulations to treat online social fragmentation as a byproduct of biased information processing and insufficient coverage due to lopsided intranetwork connectivity has been noted. However, this perception presents a conceptual challenge as it neglects to clearly distinguish between the two aspects, leading to an opaque cause-and-effect structure. Several studies have tried to determine an individual's position within an attitude space by accumulating opinions from their local network. Yet, the question remains whether biased information processing is the result or the cause of limited epistemic breadth. More critically, the role of technology, which significantly shapes the interaction process, is not sufficiently addressed by existing models that primarily depict interaction as a process of social influence.

Our findings buttress the above argument by elucidating the crucial role of technological filters in information dissemination across social networks. However, the efficacy of diversification efforts is also contingent upon the method of operationalizing individual human integration procedures. Intriguingly, in both the homophilic and ideological scenarios, the extent of information diffusion is quite similar when the technological filter limits itself to local user neighborhoods, as in *ego-network*. The difference between the two scenarios becomes noticeable only in the context of

more diversified recommendation sets.

The results suggest that *epistemic bubbles* [Ngu20] may not be as detrimental as commonly assumed. The barriers of such chambers can be breached even with *randomly* selected users and tweets. This aligns with the literature that proposes the primary cause of epistemic impairments as the mere lack of intranetwork connectivity. Conversely, the ideological scenario, characterized by preemptive distrust towards external sources, presents a greater challenge in community bridging. Even the recommendation of outsider information tailored to personal preferences (*community-intersection*) does not noticeably increase acceptance. This reflects the key characteristic of ideological echo chambers, where community insiders are shielded from potentially valid counterarguments originating from the outside.

Moreover, the outcomes related to the *boundary* and *boundary-translation* conditions indicate that border regions between communities can play a pivotal role in gradually broadening the discussion. When faced with the same argument presented by two different individuals, one from their own community and the other from an external community, acceptance is notably higher when the argument is aligned with the source within their own community. Thus, the cognitive complexity inherent in socio-digital processes must be acknowledged. As Settle's lab experiments show, social media usage intensifies perceived differences between an individual's own position and their assumed position of the out-group [Set18]. This underscores the fundamental role of social identity schemata in processing information in online environments.

3.6 Conclusion

In the rapidly evolving domain of online communication, understanding the dynamics of social interactions is both pressing and complex. Much of the existing research has been rooted in actor-theory, which predominantly emphasizes individual behaviors. However, this chapter has taken a departure from this micro-level focus, directing attention to emergent behaviors and the overarching group-level dynamics that shape online discourse.

Central to our exploration is the nuanced understanding of group identity and differentiation. Anthropological studies have illuminated the human propensity for complementary schismogenesis, a phenomenon where societies often delineate themselves not by shared attributes, but by what they distinctively reject. This inclination, anchored in group psychology, highlights the pivotal role of in-group and out-group dynamics in shaping social interactions. While these dynamics offer a heuristic to navigate social landscapes, they can, at times, result in pronounced divisions, reminiscent of historical instances such as the propaganda mechanisms in Nazi Germany.

This theoretical foundation has served to make the mechanisms of social polarization in online contexts more tangible. In particular, the realization that group-dependent processes of delimitation are an inherent mode of human world orientation has led us to reevaluate the phenomenon of digital echo chambers. Due to their elusiveness, these echo chambers pose a particular challenge both in terms of their definition and their representation in the context of simulation models, and require a more holistic perspective. Contrary to the prevailing mainstream and academic narrative that depicts them as isolated extremes characterized by communicative decoupling,

recent empirical findings suggest a more nuanced reality [Ngu20; Tö22]. Rather, ideological demarcation is a generally observable phenomenon that manifests itself in the cognitive biases and group dynamics we all use as humans to navigate our information-dense digital world. With this perspective, this chapter has delved deeper, exploring the psychological and sociological roots of these dynamics and advocating for a more nuanced understanding of social polarization, particularly the role of echo chambers in shaping online discourse.

Transitioning to concepts of social systems theory has allowed us to see the online world not as a mere collection of communicatively isolated entities but as a vast, interconnected web of communicative elements where collective influences often overshadow the individual. Echo chambers, hence better characterized as specific types of ideological social systems, are subject to the complex interplay of individual beliefs, group dynamics, and external influences. Within these systems, feedback loops play a pivotal role, amplifying certain narratives while muting others, leading to the emergence and dominance of specific ideologies. This interconnectedness suggests that our shared beliefs and communication patterns are by no means merely simple aggregates of individual views. Instead, they emerge from the collective dynamics that drive ideological system reproduction. Through the lens of systems theory, we have gained insights into these intricate feedback mechanisms, understanding how even subtle shifts can set off significant ideological transformations within online communities.

The empirical validation presented in this chapter reinforces these theoretical constructs. Patterns observed in data analysis and simulation models resonate with the systemic predictions, underscoring the broader implications of the findings. In an era dominated by online interactions, the insights derived from a systems theory perspective become invaluable, offering a roadmap to navigate the complex terrain of online ideological dynamics.

Moreover, systems theory has allowed us to consider recommender systems as actively shaping the narratives and interactions within online communities, far from being passive entities. Their influence goes beyond simply recommending content, and they play a central role in the formation and evolution of ideological groups. Our simulations have shown that despite their utility, recommender systems can inadvertently perpetuate social divisions by oversimplifying the diversity and disagreements among users. By examining the relationships between users and the messages they exchange within the network, we have sought to cultivate a more nuanced understanding of how these technological systems influence the formation of ideological communities. In addition, our examination of various recommendation technologies has provided important insights into the central role these systems play in shaping ideological landscapes online. The effectiveness of attempts to diversify information and the formation of ideological echo chambers are highly dependent on the operationalization of these technological filters. The study shows that even simple technological interventions, such as presenting randomly selected users and tweets, can help disrupt the formation of epistemic gaps, supporting the hypothesis that limited intra-network connectivity can significantly contribute to epistemic impairments. However, the task of bridging communities becomes particularly challenging in the face of the preemptive distrust of outside sources that is common in highly ideological environments.

The findings further emphasize the potential for border regions between communities to open up the discussion progressively. Understanding the nuanced interactions between users, their messages, and the recommender systems that connect them, we can better comprehend how information spreads and how opinion dynamics are affected. This understanding prompts a more

in-depth reflection on the role of recommender systems in perpetuating or mitigating ideological polarization, a direction that future research might find worthy of exploration.

In summation, this chapter offers a comprehensive exploration of the dynamics of online social systems. By bridging the domains of recommender systems, social networks, and ideological echo chambers, and through the lens of agent-based modeling, social judgment theory, and systems theory, we have provided insights into the complex interplay of technology, group psychology, and social identity in shaping online communities.

As our research advances, we identify several promising pathways for future investigation and enhancement. First and foremost, a key direction is the explicit incorporation of language into our model. In the current study, we primarily focused on the relationships between users and the messages exchanged within the network, but a deeper dive into the actual content of these messages could provide greater insight. Large Language Models, with their ability to understand and generate human-like text, offer a unique opportunity in this regard. By adopting the role of agents in our agent-based model, these models can facilitate a much more sophisticated representation of agents, enhancing the depth and realism of our simulations.

The recent open-sourcing of Twitter's algorithm also presents an exciting opportunity [Twi23]. The implementation of this algorithm within our framework allows us to approximate the dynamics of a real-world social media platform more precisely, enhancing the validity and applicability of our findings. By simulating the same algorithmic forces that shape real-world interactions on Twitter, we can gain a deeper understanding of how such forces contribute to the formation of ideological communities and the polarization of online discourse.

In addition, we foresee the development of intervention strategies aimed at mitigating the negative effects of recommender systems on social polarization and ideological isolation. As our study illustrates, recommender systems can inadvertently contribute to the formation of echo chambers and filter bubbles, leading to a skewed perception of reality and increased social division. Future research could explore how these systems might be redesigned to promote more diverse and balanced information exposure, while still catering to individual preferences and interests.

An exploration of the role of social network platform design and user interface elements in shaping communication behaviors and the formation of ideological communities is also imperative. As online social networks become increasingly pivotal to information dissemination and social redundancy creation, understanding how different platform features and affordances influence user behavior is crucial. This could provide valuable insights into how platform design might be leveraged to encourage more open and inclusive communication, counteracting the negative effects of echo chambers and filter bubbles.

Lastly, this study prompts important discussions about the ethical responsibilities of platform operators and developers of recommender systems. As we have noted, economic interests often override the need for balanced information access, leading to prioritized communication that may not be in the best interest of individuals or society. Future research could probe into the ethical implications of recommender systems, exploring the potential for regulatory measures or industry standards to ensure these technologies are developed and deployed responsibly and beneficially.

In conclusion, this thesis has shed light on the intricate dynamics underpinning the formation of ideological communities in online social networks and the role of recommender systems in these processes. As challenges posed by echo chambers, filter bubbles, and social polarization continue to call for attention, this work should serve as a springboard to delve deeper into these phenomena and to identify potential solutions fostering greater social cohesion and understanding.

Contributions

Sequential User-based Recurrent Neural Network Recommendations

In this study, we present a pioneering exploration into the utilization of deep learning techniques for the realm of recommender systems. We introduce an advanced method for generating personalized recommendations by augmenting *recurrent neural networks* [RHW86; HS97] to encapsulate the distinct attributes of the recommender systems domain. Recognizing the inherent time-dependent nature of many recommendation scenarios, our approach emphasizes the importance of time as a pivotal factor in the recommendation process. Specifically, we introduce a novel type of *gated recurrent unit* [Cho*14] that seamlessly integrates individual user representations alongside sequences of consumed items.

Furthermore, we delve into the advantages of deeply embedding user representations within the network. This profound integration amplifies the prediction of intricate behavioral patterns, such as movie-watching or music-listening tendencies. Our network is adept at explicitly representing overarching user concepts, transcending the implicit representations typically found in standard gated units. As a result, our approach can discern with greater precision between general and situational user inclinations.

We propose three distinct methods of user integration:

- *Linear user integration* views a latent user representation as an additional input layer connected to the gated units. It is comparable to strategies that consider context for gated units, such as incorporating word-related topic vectors into a neural network language model. The user component represents stable concepts, distinguishing our approach from other deep learning methods in recommender systems that often focus on consumption sessions without explicitly representing the user.
- *Rectified linear user integration* [cf. NH10] is designed to control the influence of the user component, not turning it off completely but limiting its throughput. A leaky variant is integrated, allowing the network to control information flow more accurately. The rectified linear user integration impedes the user component when necessary, providing a balance between user and item aspects.

- *Attentional user integration* employs an attention mechanism [BCB14], allowing the network to adaptively shift focus between user and item aspects in an even more dynamic manner than the rectified linear variant. The attentional gate regulates the extent to which users are considered as opposed to items, making it possible for the network to weigh the importance of different items in a user’s history and provide more context-aware recommendations.

These innovative couplings of conventional latent representations with our user-centric representations allow for a more nuanced modeling of user behavior, capturing dependencies between consumption events. This stands in stark contrast to traditional recommendation techniques that often falter in representing the evolving temporal dynamics of user interests.

This study underscores our contributions to the field by showcasing the potential of recurrent neural networks and deep learning techniques in enhancing recommendation quality. It serves as a reflection of exploring innovative approaches within the recommender systems domain, with an eye towards further refinement and understanding.

Interactive Recommending with Tag-Enhanced Matrix Factorization (TagMF)

The primary contribution of this work is the development and implementation of TagMF, a novel approach that combines user-generated tags with *matrix factorization* [KBV09] to enhance the recommendation process. This approach allows for a more interactive and personalized recommendation experience, addressing the limitations of traditional *collaborative filtering* systems [SK09] that often lack transparency and user control. A machine learning method is introduced to automatically learn relationships between latent factors and user-generated tags, which can be used, for example, to visualize a user’s long-term preference profile, which is typically non-transparent in model-based collaborative filtering.

The paper also presents an empirical user study aimed at examining the influence of additional information on decision-making in recommendation contexts and evaluating the interactive features enabled by TagMF. The study focuses in particular on validating the application possibilities of TagMF, such as using user-generated tags to elicit preferences at cold-start and interactively manipulating recommendations based on an existing user profile. Several hypotheses related to the impact of TagMF on the user experience are proposed, including improving the perceived quality of recommendations, satisfaction with the chosen item, and transparency, especially in cold-start situations. The study also hypothesizes that TagMF decreases the difficulty of choosing an item and does not negatively impact perceived interaction effort.

Explaining Recommendations by Means of User Reviews

This positional paper introduces a conceptual framework for leveraging user-generated content, such as product reviews, to enhance the transparency and user control of recommender systems. We contend that despite significant advancements in algorithmic sophistication and recommendation quality, recommender systems often operate as black boxes, offering little explanation for their recommendations.

To address this issue, we propose an innovative approach that extracts complex argumentative explanations from user reviews and presents them to users. This approach comprises two main

components: linguistic analyses for *argument mining* [PM09; LT16] and *stance detection* [KC20], and an *attention-based mechanism* [BCB14] for identifying important concepts for a target user. The linguistic analyses are designed to derive argumentative structures from reviews, including argument polarity in the form of stances. Conversely, the attention-based mechanism aims to mimic human scanning behavior by adaptively assigning individual attention weights to arguments based on a user's preferences.

We also suggest a method for deriving an argumentative flow through multiple applications of the attention-based mechanism. This results in a succession of relevant word sets that exhibit sequential properties and are interconnected. This process allows the system to lay out the entire path of how it arrived at a particular recommendation.

Finally, we discuss the challenges and future research directions in this field, including the implementation of this novel concept, its use for personalizing recommendations, and its evaluation in various domains. We particularly focus on its influence on user experience compared to other explanation methods.

Explaining Recommendations by Means of Aspect-based Transparent Memories

Building on the conceptual framework presented in our previous work, this research introduces a novel approach to recommendation systems that harnesses both explicit and implicit user knowledge to generate aspect-based justifications for recommendations. This concrete implementation and evaluation of our earlier positional paper underscores the significance of user studies in recommendation systems and research on explainable AI, offering valuable insights into the generation of justifications.

Our primary contribution is the development of Aspect-based Transparent Memories (ATM), a *neural memory-based* [WCB15; Suk*15] recommendation approach designed to generate aspect-based justifications. This method leverages both explicit and implicit knowledge about the user, providing a comprehensive representation of the user's preferences. The explicit knowledge is derived from user reviews, while the implicit knowledge captures latent patterns in a person's rating behavior, akin to conventional collaborative filtering.

We also introduce an attention mechanism to distinguish between relevant and irrelevant sections of a review, based on the assumption that people express thoughts and opinions in more or less fixed linguistic units. This mechanism allows for the distribution of attention weights at the sentence level, offering a more granular understanding of user preferences.

Beyond the technical contributions, we present an extensive user study aimed at evaluating the quality of justifications as assessed by people. The study uncovers three central interacting factors contributing to the quality of textual justifications: content adequacy, presentation adequacy, and linguistic adequacy. The results of the study are analyzed using robust statistical tools, providing valuable insights into the generation of effective textual justifications in a decision task.

We also highlight the general importance of user studies in recommendation systems and explainable AI research. We argue that offline evaluations often fail to test whether justifications truly assist people in making decisions. The user study conducted in our work provides insights

into the factors that contribute to the generation of effective textual justifications and how these factors interact with each other.

Leveraging Arguments in User Reviews for Generating and Explaining Recommendations

This paper offers another take on the ATM approach from a different perspective, focusing on the application of argumentation theory in the context of recommendation systems. It explores the idea of viewing recommendations as a form of argumentation, where a recommendation is a claim that the user will find the recommended item useful or pleasing. It emphasizes that recommendations do not aim at influencing a person's long-term beliefs, but rather at supporting users in their decision-making in a specific interactive context, thus involving a strongly persuasive component.

Furthermore, the paper discusses the limitations of conventional recommender systems, which often function as black boxes and do not provide the user with explanations for a given recommendation. It argues that the relationship between a recommendation and its explanation can be considered a specific form of argumentation, although very little research has thus far investigated explainability from this perspective.

The paper also highlights the potential of ATM in generating justifications for recommendations. It points out that while ATM is mostly concerned with detecting personally important information in review texts composed by other users, the extracted content is not yet presented in an argumentative manner as no structural knowledge about arguments is represented in the model. This leads to several limitations, such as the fact that the resulting explanations only express merely apparent causality, motivating the application of causal reasoning techniques in the future.

The Dual Echo Chamber: Modeling Social Media Polarization for Interventional Recommending

The paper presents a novel approach to understanding and addressing the issue of polarization in social media platforms. The main contribution of this work lies in the development of an *agent-based modeling* algorithm [RG19; EA96] that simulates user interaction with recommendations generated by a *knowledge graph embedding* model [Bor*13; Wan*17; Ham*18].

The algorithm is designed to model echo chambers as a process of local information distribution. It uses a set of users and messages, and based on the edges until the current temporal bin, it learns to embed relations between these entities as deductive or inductive knowledge search. The algorithm also integrates a mapping between communities and users, allowing the model to query the user space from communities, effectively identifying representative users of a community.

The paper also introduces the concept of *geometric operators* [Ham*18] that are translated into differentiable form and optimized along the embeddings for graph nodes. These operators include the geometric translation operator and the geometric intersection operator. The former outputs a new query embedding given a query embedding and an edge type, while the latter performs set intersection in embedding space.

The study also includes a detailed evaluation protocol. It verifies the effectiveness of the proposed approach in terms of link prediction quality and presents the results of the agent-based modeling procedure in detail. The results show that the proposed method consistently outperforms the baseline, indicating its effectiveness in addressing the issue of social media polarization.

Finally, the paper discusses the implementation of cognitive filters, i.e., individual and ideological demarcation, in the modeling process. It highlights that while both scenarios follow comparable dynamics when modeling echo chambers, notably different patterns can be observed when recommendations are diversified. This stresses the importance of evaluating the effectiveness of depolarization techniques with respect to carefully designed human decision procedures.

De-sounding Echo Chambers: Simulation-based Analysis of Polarization Dynamics in Social Networks

This work establishes an extensive framework, based on the previous work, that underscores the interaction between human decision-making processes and the impact of recommendation technology on network dynamics. This interaction is pivotal for understanding the dynamics of social polarization.

The cognitive biases leading individuals to engage with those who share similar views, while distancing themselves from unfamiliar or novel ideas, are acknowledged. This selective exposure to information can, however, only sustain a social divide temporarily if certain network structures inhibit the emergence of a shared consensus.

Beyond selective exposure, the exploration of other psychological mechanisms that contribute to the formation of echo chambers is a significant aspect of the work. Systematic alienation from external knowledge sources, which discourages individuals from seeking information outside their intellectual community and motivates them to actively discredit external voices, is highlighted. This systematic exclusion of information sources driven by ideological demarcation is posited as a key factor in polarization. In this context, the differentiation between in-group and out-group, a fundamental element of social organization, is examined. Our findings suggest that compartmentalization can be identified not only at the extreme edges of a debate but also among representatives of mainstream viewpoints.

The primary focus of the work is to demonstrate the ability to model polarization in various scenarios, ranging from binary to multi-group polarization. Each scenario presents different degrees of polarization and inter-group overlap, especially concerning the initial starting graph. Visualizations of graph topologies and latent space polarization for various scenarios are provided, illustrating how different conditions using ideological demarcation can separate communities to varying degrees.

The evolution of polarization over time, revealing how different conditions affect this process, is also presented. Various metrics are calculated to analyze network and community structure and quantify polarization between communities. However, the findings acknowledge that some structural details are lost in certain conditions, which motivates the future application of hierarchical community structure to represent sub-community structures.

Bibliography

- [Ado13] Gediminas Adomavicius. “Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects.” In: *eBusiness & eCommerce eJournal* (2013).
- [AH97] Theodor W Adorno and Max Horkheimer. *Dialectic of enlightenment*. Vol. 15. Verso, 1997.
- [Ado70] W Adorno Theodor. “Die Freudsche Theorie und die Struktur der faschistischen Propaganda.” In: *Psyche* 24.7 (1970). Publisher: Berlin: Kemper, pp. 486–509.
- [AC09] Deepak Agarwal and Bee-Chung Chen. “Regression-based latent factor models.” In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009, pp. 19–28.
- [AAJ22] Sajad Ahmadian, Milad Ahmadian, and Mahdi Jalili. “A deep learning based trust-and tag-aware recommender system.” In: *Neurocomputing* 488 (2022). Publisher: Elsevier, pp. 557–571.
- [Are73] Hannah Arendt. *The origins of totalitarianism*. Vol. 244. Houghton Mifflin Harcourt, 1973.
- [Arr*22] Henrique Ferraz de Arruda, Felipe Maciel Cardoso, Guilherme Ferraz de Arruda, Alexis R Hernández, Luciano da Fontoura Costa, and Yamir Moreno. “Modelling how social network algorithms can influence opinion polarization.” In: *Information Sciences* 588 (2022). Publisher: Elsevier, pp. 265–278.
- [BM57] Francis Bacon and Basil Montagu. *The Works of Francis Bacon*. Vol. 1. Parry & McMillan, 1857.
- [BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate.” In: *arXiv preprint arXiv:1409.0473* (2014).
- [BMA15] Eytan Bakshy, Solomon Messing, and Lada A Adamic. “Exposure to ideologically diverse news and opinion on Facebook.” In: *Science* 348.6239 (2015). Publisher: American Association for the Advancement of Science, pp. 1130–1132.
- [Bar*15] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. “Tweeting from left to right: Is online political communication more than an echo chamber?” In: *Psychological science* 26.10 (2015). Publisher: Sage Publications Sage CA: Los Angeles, CA, pp. 1531–1542.
- [Bar79] Roland Barthes. *Mythologies*. Paris: Editions du Seuil, 1979. ISBN: 2-02-000585-9.
- [Bat58] Gregory Bateson. *Naven: A survey of the problems suggested by a composite picture of the culture of a New Guinea tribe drawn from three points of view*. Vol. 21. Stanford University Press, 1958.
- [Bau*20] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. “Modeling echo chambers and polarization dynamics in social networks.” In: *Physical Review Letters* 124.4 (2020). Publisher: APS, p. 048301.

- [Bee*16] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. "Paper recommender systems: a literature survey." In: *International Journal on Digital Libraries* 17 (2016). Publisher: Springer, pp. 305–338.
- [BL15] Joeran Beel and Stefan Langer. "A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems." In: *Research and Advanced Technology for Digital Libraries*. Ed. by Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla. Cham: Springer International Publishing, 2015, pp. 153–168. ISBN: 978-3-319-24592-8.
- [BK07] Robert M Bell and Yehuda Koren. "Lessons from the netflix prize challenge." In: *Acm Sigkdd Explorations Newsletter* 9.2 (2007). Publisher: ACM New York, NY, USA, pp. 75–79.
- [BM*11] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. "The Million Song Dataset." In: *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*. 2011.
- [Bey94] Matthias Beyerle. "Staatstheorie und Autopoiesis." In: *Über die Auflösung der modernen Staatsidee im nachmodernen Denken durch die Theorie autopoietischer Systeme und der Entwurf eines nachmodernes Staatskonzepts*. Frankfurt, aM: Peter Lang (1994).
- [Bol*10] Dirk Bollen, Bart P Knijnenburg, Martijn C Willemsen, and Mark Graus. "Understanding choice overload in recommender systems." In: *Proceedings of the fourth ACM conference on Recommender systems*. 2010, pp. 63–70.
- [Bor*13] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. "Translating embeddings for modeling multi-relational data." In: *Advances in neural information processing systems* 26 (2013).
- [Bor*09] Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. "Network analysis in the social sciences." In: *science* 323.5916 (2009). Publisher: American Association for the Advancement of Science, pp. 892–895.
- [Bor03] Gary Bornstein. "Intergroup conflict: Individual, group, and collective interests." In: *Personality and social psychology review* 7.2 (2003). Publisher: Sage Publications Sage CA: Los Angeles, CA, pp. 129–145.
- [BOH12] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. "TasteWeights: a visual interactive hybrid recommender system." In: *Proceedings of the sixth ACM conference on Recommender systems*. 2012, pp. 35–42.
- [Bra97] Dietmar Braun. "Politische Steuerung zwischen Akteurs-und Systemtheorie." In: *Politische Vierteljahresschrift* 38.4 (1997), pp. 844–854.
- [Bri*20] Jonathan Bright, Nahema Marchal, Bharath Ganesh, and Stevan Rudinac. "Echo chambers exist! (But they're full of opposing views)." In: *arXiv preprint arXiv:2001.11461* (2020).
- [Bro*20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. "Language models are few-shot learners." In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.
- [CB07] Michel Callon and Oxford Blackwell. "Actor-network theory." In: *The Politics of Interventions, Oslo Academic Press, Unipub, Oslo* 1 (2007), pp. 273–286.

- [CHV22] Pablo Castells, Neil Hurley, and Saúl Vargas. “Novelty and Diversity in Recommender Systems.” In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, and Bracha Shapira. New York, NY: Springer US, 2022, pp. 603–646. ISBN: 978-1-07-162197-4. DOI: 10.1007/978-1-0716-2197-4_16.
- [Cha*14] HF Chau, CY Wong, FK Chow, and Chi-Hang Fred Fung. “Social judgment theory based model on opinion formation, polarization and evolution.” In: *Physica A: Statistical Mechanics and its Applications* 415 (2014). Publisher: Elsevier, pp. 133–140.
- [Che*23] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. “Bias and debias in recommender system: A survey and future directions.” In: *ACM Transactions on Information Systems* 41.3 (2023). Publisher: ACM New York, NY, pp. 1–39.
- [CCW15] Li Chen, Guanliang Chen, and Feng Wang. “Recommender systems based on user reviews: the state of the art.” In: *User Modeling and User-Adapted Interaction* 25 (2015). Publisher: Springer, pp. 99–154.
- [CP12] Li Chen and Pearl Pu. “Critiquing-based recommenders: survey and emerging trends.” In: *User Modeling and User-Adapted Interaction* 22 (2012). Publisher: Springer, pp. 125–150.
- [Che*19] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. “How serendipity improves user satisfaction with recommendations? a large-scale user evaluation.” In: *The world wide web conference*. 2019, pp. 240–250.
- [CZW22] Xu Chen, Yongfeng Zhang, and Ji-Rong Wen. “Measuring” Why” in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation.” In: *arXiv preprint arXiv:2202.06466* (2022).
- [CM06] Judith A Chevalier and Dina Mayzlin. “The effect of word of mouth on sales: Online book reviews.” In: *Journal of marketing research* 43.3 (2006). Publisher: SAGE Publications Sage CA: Los Angeles, CA, pp. 345–354.
- [Cho*14] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation.” In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179.
- [CVB14] Mina Cikara and Jay J Van Bavel. “The neuroscience of intergroup relations: An integrative review.” In: *Perspectives on Psychological Science* 9.3 (2014). Publisher: Sage Publications Sage CA: Los Angeles, CA, pp. 245–274.
- [CAS16] Paul Covington, Jay Adams, and Emre Sargin. “Deep Neural Networks for Youtube Recommendations.” In: *Proceedings of the 10th ACM conference on recommender systems*. 2016, pp. 191–198.
- [DGL13] Pranav Dandekar, Ashish Goel, and David T Lee. “Biased assimilation, homophily, and the dynamics of polarization.” In: *Proceedings of the National Academy of Sciences* 110.15 (2013). Publisher: National Acad Sciences, pp. 5791–5796.

- [DGM14] Abhimanyu Das, Sreenivas Gollapudi, and Kamesh Munagala. “Modeling opinion dynamics in social networks.” In: *Proceedings of the 7th ACM international conference on Web search and data mining*. 2014, pp. 403–412.
- [DG*15] Marco De Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. “Semantics-aware content-based recommender systems.” In: *Recommender systems handbook (2015)*. Publisher: Springer, pp. 119–159.
- [Dec*22] Oliver Decker, Johannes Kiess, Aylene Heller, Julia Schuler, and Elmar Brähler. “Die Leipziger Autoritarismus Studie 2022: Methode, Ergebnisse und Langzeitverlauf.” In: *Autoritäre Dynamiken in unsicheren Zeiten*. Psychosozial-Verlag, 2022, pp. 31–90.
- [Def*00] Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. “Mixing beliefs among interacting agents.” In: *Advances in Complex Systems* 3.01n04 (2000), pp. 87–98.
- [DeG74] Morris H DeGroot. “Reaching a consensus.” In: *Journal of the American Statistical association* 69.345 (1974). Publisher: Taylor & Francis, pp. 118–121.
- [DV*15] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. “Echo chambers in the age of misinformation.” In: *arXiv preprint arXiv:1509.00189* (2015).
- [Del*23] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. “Fairness in recommender systems: research landscape and future directions.” In: *User Modeling and User-Adapted Interaction (2023)*. Publisher: Springer, pp. 1–50.
- [DLZ17] Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. “Sequential User-Based Recurrent Neural Network Recommendations.” In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. RecSys ’17. event-place: Como, Italy. New York, NY, USA: Association for Computing Machinery, 2017, pp. 152–160. ISBN: 978-1-4503-4652-8. DOI: 10.1145/3109859.3109877.
- [DKZ20] Tim Donkers, Timm Kleemann, and Jürgen Ziegler. “Explaining recommendations by means of aspect-based transparent memories.” In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 2020, pp. 166–176.
- [DZ21] Tim Donkers and Jürgen Ziegler. “The dual echo chamber: Modeling social media polarization for interventional recommending.” In: *Proceedings of the 15th ACM Conference on Recommender Systems*. 2021, pp. 12–22.
- [DZ23] Tim Donkers and Jürgen Ziegler. “De-sounding echo chambers: Simulation-based analysis of polarization dynamics in social networks.” In: *Online Social Networks and Media* 37-38 (2023), p. 100275. ISSN: 2468-6964. DOI: <https://doi.org/10.1016/j.osnem.2023.100275>.
- [Dou*19] Karen M Douglas, Joseph E Uscinski, Robbie M Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. “Understanding conspiracy theories.” In: *Political psychology* 40 (2019). Publisher: Wiley Online Library, pp. 3–35.
- [Eks*22] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, and others. “Fairness in information access systems.” In: *Foundations and Trends® in Information Retrieval* 16.1-2 (2022). Publisher: Now Publishers, Inc., pp. 1–177.

- [EA96] Joshua M Epstein and Robert Axtell. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press, 1996.
- [FH09a] Armin Falk and James J Heckman. “Lab experiments are a major source of knowledge in the social sciences.” In: *science* 326.5952 (2009). Publisher: American Association for the Advancement of Science, pp. 535–538.
- [Faz18] Soude Fazeli. “User-Centric Evaluation of Recommender Systems in Social Learning Platforms: Accuracy is Just the Tip of the Iceberg.” In: *IEEE Transactions on Learning Technologies* (2018).
- [Fes62] Leon Festinger. “Cognitive dissonance.” In: *Scientific American* 207.4 (1962). Publisher: JSTOR, pp. 93–106.
- [FTA22] Giacomo Figà Talamanca and Selene Arfini. “Through the newsfeed glass: Rethinking filter bubbles and Echo chambers.” In: *Philosophy & Technology* 35.1 (2022). Publisher: Springer, p. 20.
- [FGR16] Seth Flaxman, Sharad Goel, and Justin M Rao. “Filter bubbles, echo chambers, and online news consumption.” In: *Public opinion quarterly* 80.S1 (2016). Publisher: Oxford University Press, pp. 298–320.
- [FH09b] Daniel Fleder and Kartik Hosanagar. “Blockbuster culture’s next rise or fall: The impact of recommender systems on sales diversity.” In: *Management science* 55.5 (2009). Publisher: INFORMS, pp. 697–712.
- [FZ11] Peter Forbes and Mu Zhu. “Content-boosted matrix factorization for recommender systems: experiments with recipe recommendation.” In: *Proceedings of the fifth ACM conference on Recommender systems*. 2011, pp. 261–264.
- [Fre89] Sigmund Freud. *Group psychology and the analysis of the ego*. WW Norton & Company, 1989.
- [Fre86] Dieter Frey. “Recent research on selective exposure to information.” In: *Advances in experimental social psychology* 19 (1986). Publisher: Elsevier, pp. 41–80.
- [Fro21] Erich Fromm. *The fear of freedom*. Routledge, 2021.
- [Gar23] Eric Garcia. “GOP uses dystopian AI-generated video to attack Biden as he launches re-election bid.” In: (April 2023). URL: <https://www.independent.co.uk/news/world/americas/us-politics/republican-biden-2024-reelection-ai-b2326586.html> (visited on October 16, 2023).
- [GDBJ10] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. “Beyond accuracy: evaluating recommender systems by coverage and serendipity.” In: *Proceedings of the fourth ACM conference on Recommender systems*. 2010, pp. 257–260.
- [GJG14] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. “How should I explain? A comparison of different explanation types for recommender systems.” In: *International Journal of Human-Computer Studies* 72.4 (2014). Publisher: Elsevier, pp. 367–382.
- [GM21] Negin Ghasemi and Saeedeh Momtazi. “Neural text similarity of user reviews for improving collaborative filtering recommender systems.” In: *Electronic Commerce Research and Applications* 45 (2021). Publisher: Elsevier, p. 101019.
- [GVL13] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

- [GUH15] Carlos A Gomez-Urbe and Neil Hunt. “The netflix recommender system: Algorithms, business value, and innovation.” In: *ACM Transactions on Management Information Systems (TMIS)* 6.4 (2015). Publisher: ACM New York, NY, USA, pp. 1–19.
- [Goo*20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative adversarial networks.” In: *Communications of the ACM* 63.11 (2020). Publisher: ACM New York, NY, USA, pp. 139–144.
- [GJ17] Jyotirmoy Gope and Sanjay Kumar Jain. “A survey on solving cold start problem in recommender systems.” In: *2017 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 2017, pp. 133–138.
- [GW21] David Graeber and David Wengrow. *The dawn of everything: A new history of humanity*. Penguin UK, 2021.
- [GC17] Damaris Graeupner and Alin Coman. “The dark side of meaning-making: How social exclusion leads to superstitious thinking.” In: *Journal of Experimental Social Psychology* 69 (2017). Publisher: Elsevier, pp. 218–222.
- [Gue*18] Andrew Guess, Brendan Nyhan, Benjamin Lyons, and Jason Reifler. “Avoiding the echo chamber about echo chambers.” In: *Knight Foundation* 2 (2018), pp. 1–25.
- [Gö02] Axel Görlitz. “Politische Steuerung.” In: *Handwörterbuch zur politischen Kultur der Bundesrepublik Deutschland* (2002). Publisher: Springer, pp. 459–467.
- [HG17] Ivan Habernal and Iryna Gurevych. “Argumentation Mining in User-Generated Web Discourse.” In: *Computational Linguistics* 43.1 (April 2017). Place: Cambridge, MA Publisher: MIT Press, pp. 125–179. DOI: 10.1162/COLI_a_00276.
- [Ham*18] Will Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. “Embedding logical queries on knowledge graphs.” In: *Advances in neural information processing systems* 31 (2018).
- [Har*14] Jason L Harman, John O’Donovan, Tarek Abdelzaher, and Cleotilde Gonzalez. “Dynamics of human trust in recommender systems.” In: *Proceedings of the 8th ACM Conference on Recommender systems*. 2014, pp. 305–308.
- [HK15] F Maxwell Harper and Joseph A Konstan. “The movielens datasets: History and context.” In: *Acm transactions on interactive intelligent systems (tiis)* 5.4 (2015). Publisher: Acm New York, NY, USA, pp. 1–19.
- [HPV16] Chen He, Denis Parra, and Katrien Verbert. “Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities.” In: *Expert Systems with Applications* 56 (2016). Publisher: Elsevier, pp. 9–27.
- [HM16] Ruining He and Julian McAuley. “Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering.” In: *proceedings of the 25th international conference on world wide web*. 2016, pp. 507–517.
- [HK02] Rainer Hegselmann and Ulrich Krause. “Opinion Dynamics and Bounded Confidence: Models, Analysis and Simulation.” In: *Journal of Artificial Societies and Social Simulation* 5.3 (2002).
- [HBZ23] Diana C Hernandez-Bocanegra and Jürgen Ziegler. “Explaining Recommendations through Conversations: Dialog Model and the Effects of Interface Type and Degree

- of Interactivity.” In: *ACM Transactions on Interactive Intelligent Systems* 13.2 (2023). Publisher: ACM New York, NY, pp. 1–47.
- [HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory.” In: *Neural computation* 9.8 (1997). Publisher: MIT press, pp. 1735–1780.
- [HLZ08] Nan Hu, Ling Liu, and Jie Jennifer Zhang. “Do online reviews affect product sales? The role of reviewer characteristics and temporal effects.” In: *Information Technology and management* 9 (2008). Publisher: Springer, pp. 201–214.
- [HKV08] Yifan Hu, Yehuda Koren, and Chris Volinsky. “Collaborative filtering for implicit feedback datasets.” In: *2008 Eighth IEEE international conference on data mining. Ieee*, 2008, pp. 263–272.
- [JA05] Wander Jager and Frédéric Amblard. “Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change.” In: *Computational & Mathematical Organization Theory* 10.4 (2005). Publisher: Springer, pp. 295–303.
- [Jia*19] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. “Degenerate feedback loops in recommender systems.” In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 383–390.
- [KKH15] Ece Kamar, Ashish Kapoor, and Eric Horvitz. “Identifying and accounting for task-dependent bias in crowdsourcing.” In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 3. 2015, pp. 92–101.
- [KB16] Marius Kaminskis and Derek Bridge. “Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems.” In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7.1 (2016). Publisher: ACM New York, NY, USA, pp. 1–42.
- [Kar*17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. “Progressive growing of gans for improved quality, stability, and variation.” In: *arXiv preprint arXiv:1710.10196* (2017).
- [Kla95] Joshua Klayman. “Varieties of confirmation bias.” In: *Psychology of learning and motivation* 32 (1995). Publisher: Elsevier, pp. 385–418.
- [Kni*12] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. “Explaining the user experience of recommender systems.” In: *User modeling and user-adapted interaction* 22 (2012). Publisher: Springer, pp. 441–504.
- [KR12] Joseph A Konstan and John Riedl. “Recommender systems: from algorithms to user experience.” In: *User modeling and user-adapted interaction* 22 (2012). Publisher: Springer, pp. 101–123.
- [KBV09] Yehuda Koren, Robert Bell, and Chris Volinsky. “Matrix factorization techniques for recommender systems.” In: *Computer* 42.8 (2009). Publisher: IEEE, pp. 30–37.
- [KRB21] Yehuda Koren, Steffen Rendle, and Robert Bell. “Advances in Collaborative Filtering.” In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, and Bracha Shapira. New York, NY: Springer US, 2021, pp. 91–142. ISBN: 978-1-07-162197-4. DOI: 10.1007/978-1-0716-2197-4_3.

- [KWV16] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. “A survey of serendipity in recommender systems.” In: *Knowledge-Based Systems* 111 (2016). Publisher: Elsevier, pp. 180–192.
- [KP17] Matevž Kunaver and Tomaž Požrl. “Diversity in recommender systems—A survey.” In: *Knowledge-based systems* 123 (2017). Publisher: Elsevier, pp. 154–162.
- [Kun*19] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. “Let me explain: Impact of personal and impersonal explanations on trust in recommender systems.” In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–12.
- [KLZ17] Johannes Kunkel, Benedikt Loepp, and Jürgen Ziegler. “A 3D item space visualization for presenting and manipulating user preferences in collaborative filtering.” In: *Proceedings of the 22nd international conference on intelligent user interfaces*. 2017, pp. 3–15.
- [KC20] Dilek Küçük and Fazli Can. “Stance detection: A survey.” In: *ACM Computing Surveys (CSUR)* 53.1 (2020). Publisher: ACM New York, NY, USA, pp. 1–37.
- [LM14] Ernesto Laclau and Chantal Mouffe. *Hegemony and socialist strategy: Towards a radical democratic politics*. Vol. 8. Verso Books, 2014.
- [LB02] Gustave Le Bon. *The crowd: A study of the popular mind*. Courier Corporation, 2002.
- [LeC*98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition.” In: *Proceedings of the IEEE* 86.11 (1998). Publisher: Ieee, pp. 2278–2324.
- [LH19] Dokyun Lee and Kartik Hosanagar. “How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment.” In: *Information Systems Research* 30.1 (2019). Publisher: INFORMS, pp. 239–259.
- [Leo21] David Leonhardt. “Covid’s Partisan Errors.” In: (March 2021). URL: <https://www.nytimes.com/2021/03/18/briefing/atlanta-shootings-kamala-harris-tax-deadline-2021.html> (visited on October 20, 2023).
- [Les*09] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. “Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters.” In: *Internet Mathematics* 6.1 (2009). Publisher: Taylor & Francis, pp. 29–123.
- [LKH14] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. “Facing the cold start problem in recommender systems.” In: *Expert systems with applications* 41.4 (2014). Publisher: Elsevier, pp. 2065–2073.
- [LT16] Marco Lippi and Paolo Torroni. “Argumentation mining: State of the art and emerging trends.” In: *ACM Transactions on Internet Technology (TOIT)* 16.2 (2016). Publisher: ACM New York, NY, USA, pp. 1–25.
- [Loe*19] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. “Interactive recommending with tag-enhanced matrix factorization (TagMF).” In: *International Journal of Human-Computer Studies* 121 (2019). Publisher: Elsevier, pp. 21–41.
- [Lov*18] Matt Lovett, Saleh Bajaba, Myra Lovett, and Marcia J Simmering. “Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of Amazon’s Me-

- chanical Turk Masters.” In: *Applied Psychology* 67.2 (2018). Publisher: Wiley Online Library, pp. 339–366.
- [LRT22] Nadia Loy, Matteo Raviola, and Andrea Tosin. “Opinion polarization in social networks.” In: *Philosophical Transactions of the Royal Society A* 380.2224 (2022). Publisher: The Royal Society, p. 20210158.
- [Luh95] Niklas Luhmann. *Social systems*. stanford university Press, 1995.
- [Luh97] Niklas Luhmann. *Die Gesellschaft der Gesellschaft*. Vol. 2. Suhrkamp Frankfurt am Main, 1997.
- [LC00] Niklas Luhmann and Kathleen Cross. *The reality of the mass media*. Stanford University Press Stanford, CA, 2000.
- [MBP18] Jens Koed Madsen, Richard M Bailey, and Toby D Pilditch. “Large networks of rational agents form persistent echo chambers.” In: *Scientific reports* 8.1 (2018). Publisher: Nature Publishing Group, pp. 1–8.
- [Mau00] Marcel Mauss. *The gift: The form and reason for exchange in archaic societies*. WW Norton & Company, 2000.
- [McS05] David McSherry. “Explanation in recommender systems.” In: *Artificial Intelligence Review* 24 (2005). Publisher: Springer, pp. 179–197.
- [Mik*13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient estimation of word representations in vector space.” In: *arXiv preprint arXiv:1301.3781* (2013).
- [MTF20] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. “Recommender systems and their ethical challenges.” In: *Ai & Society* 35 (2020). Publisher: Springer, pp. 957–967.
- [MM11] Raquel Mochales and Marie-Francine Moens. “Argumentation mining.” In: *Artificial Intelligence and Law* 19 (2011). Publisher: Springer, pp. 1–22.
- [Mol13] Pascal Molenberghs. “The neuroscience of in-group bias.” In: *Neuroscience & Biobehavioral Reviews* 37.8 (2013). Publisher: Elsevier, pp. 1530–1536.
- [MR21] Xavier Molinero and Fabián Riquelme. “Influence decision models: From cooperative game theory to social network analysis.” In: *Computer Science Review* 39 (2021). Publisher: Elsevier, p. 100343.
- [MST20] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. “Explaining machine learning classifiers through diverse counterfactual explanations.” In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 607–617.
- [Mou99] Chantal Mouffe. “Deliberative democracy or agonistic pluralism?” In: *Social research* (1999). Publisher: JSTOR, pp. 745–758.
- [Mur98] Jonathan Murdoch. “The spaces of actor-network theory.” In: *Geoforum* 29.4 (1998). Publisher: Elsevier, pp. 357–374.
- [MMT18] Cameron Musco, Christopher Musco, and Charalampos E Tsourakakis. “Minimizing polarization and disagreement in social networks.” In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 369–378.
- [NH10] Vinod Nair and Geoffrey E Hinton. “Rectified linear units improve restricted boltzmann machines.” In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 807–814.

- [NE02] Leonard S Newman and Ralph Erber. *Understanding genocide: The social psychology of the Holocaust*. Oxford University Press, 2002.
- [Ngu20] C Thi Nguyen. “Echo chambers and epistemic bubbles.” In: *Episteme* 17.2 (2020). Publisher: Cambridge University Press, pp. 141–161.
- [Ngu14] Tien T. Nguyen. “Exploring the filter bubble: the effect of using recommender systems on content diversity.” In: *Proceedings of the 23rd international conference on World wide web* (2014).
- [NN93] Elisabeth Noelle-Neumann. *The spiral of silence: Public opinion—Our social skin*. University of Chicago Press, 1993.
- [OMK20] Daniel W Otter, Julian R Medina, and Jugal K Kalita. “A survey of the usages of deep learning for natural language processing.” In: *IEEE transactions on neural networks and learning systems* 32.2 (2020). Publisher: IEEE, pp. 604–624.
- [PM09] Raquel Mochales Palau and Marie-Francine Moens. “Argumentation mining: the detection, classification and structure of arguments in text.” In: *Proceedings of the 12th international conference on artificial intelligence and law*. 2009, pp. 98–107.
- [Par11] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.
- [Pec19] Florian Pecune. “A Model of Social Explanations for a Conversational Movie Recommendation System.” In: *Proceedings of the 7th International Conference on Human-Agent Interaction* (2019).
- [PM20] Hafizh A Prasetya and Tsuyoshi Murata. “A model of opinion and propagation structure polarization in social media.” In: *Computational Social Networks* 7.1 (2020). Publisher: Springer, pp. 1–35.
- [PMC15] Anton V Proskurnikov, Alexey S Matveev, and Ming Cao. “Opinion dynamics in social networks with hostile camps: Consensus vs. polarization.” In: *IEEE Transactions on Automatic Control* 61.6 (2015). Publisher: IEEE, pp. 1524–1536.
- [PCH11] Pearl Pu, Li Chen, and Rong Hu. “A user-centric evaluation framework for recommender systems.” In: *Proceedings of the fifth ACM conference on Recommender systems*. 2011, pp. 157–164.
- [Rad*19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others. “Language models are unsupervised multitask learners.” In: *OpenAI blog* 1.8 (2019), p. 9.
- [RG19] Steven F Railsback and Volker Grimm. *Agent-based and individual-based modeling: a practical introduction*. Princeton university press, 2019.
- [RD22] Shaina Raza and Chen Ding. “News recommender system: a review of recent progress, challenges, and opportunities.” In: *Artificial Intelligence Review* (2022). Publisher: Springer, pp. 1–52.
- [RA20] Urbano Reviglio and Claudio Agosti. “Thinking outside the black-box: The case for “algorithmic sovereignty” in social media.” In: *Social Media+ Society* 6.2 (2020). Publisher: SAGE Publications Sage UK: London, England, p. 2056305120915613.
- [RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ““ Why should i trust you?” Explaining the predictions of any classifier.” In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

- [Rom*22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. “High-resolution image synthesis with latent diffusion models.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.
- [Ros73] Eleanor H Rosch. “Natural categories.” In: *Cognitive psychology* 4.3 (1973). Publisher: Elsevier, pp. 328–350.
- [RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. “Learning Internal Representations by Error Propagation.” In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362. ISBN: 0-262-68053-X.
- [Sah13] Marshall Sahlin. *Culture and practical reason*. University of Chicago Press, 2013.
- [SLL21] Fernando P Santos, Yphtach Lelkes, and Simon A Levin. “Link recommendation algorithms and dynamics of polarization in online social networks.” In: *Proceedings of the National Academy of Sciences* 118.50 (2021). Publisher: National Acad Sciences, e2102141118.
- [Sap17] Robert M Sapolsky. *Behave: The biology of humans at our best and worst*. Penguin, 2017.
- [Sas*21] Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. “Social influence and unfollowing accelerate the emergence of echo chambers.” In: *Journal of Computational Social Science* 4.1 (2021). Publisher: Springer, pp. 381–402.
- [Sau11] Ferdinand Mongin Saussure. *Course in general linguistics*. Columbia University Press, 2011.
- [Sch*07] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. “Collaborative filtering recommender systems.” In: *The adaptive web: methods and strategies of web personalization* (2007). Publisher: Springer, pp. 291–324.
- [SD16] Daan Scheepers and Belle Derks. “Revisiting social identity theory from a neuroscience perspective.” In: *Current Opinion in Psychology* 11 (2016). Publisher: Elsevier, pp. 74–78.
- [Sch08] Carl Schmitt. *The concept of the political: Expanded edition*. University of Chicago Press, 2008.
- [Sch22] Bernhard Schölkopf. “Causality for machine learning.” In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 2022, pp. 765–804.
- [Set18] Jaime E Settle. *Frenemies: How social media polarizes America*. Cambridge University Press, 2018.
- [She*61] Muzafer Sherif, O.J. Harvey, Jack B. White, William R. Hood, and Carolyn W. Sherif. *Intergroup Conflict and Cooperation. The Robbers Cave Experiment [1954]*. Norman: University of Oklahoma Book Exchange, 1961.
- [SH61] Muzafer Sherif and Carl I. Hovland. *Social judgment: Assimilation and contrast effects in communication and attitude change*. Publisher: Yale Univer. Press. 1961.
- [Sim13] Henri Simula. “The rise and fall of crowdsourcing?” In: *2013 46th Hawaii International Conference on System Sciences*. IEEE, 2013, pp. 2783–2791.
- [SL17] Brent Smith and Greg Linden. “Two decades of recommender systems at Amazon.com.” In: *IEEE internet computing* 21.3 (2017). Publisher: IEEE, pp. 12–18.

- [Son*21] Nasim Sonboli, Jessie J Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. "Fairness and transparency in recommendation: The users' perspective." In: *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 2021, pp. 274–279.
- [SB17] Hyunjin Song and Hajo G Boomgaarden. "Dynamic spirals put to test: An agent-based model of reinforcing spirals between selective exposure, interpersonal networks, and attitude polarization." In: *Journal of Communication* 67.2 (2017). Publisher: Oxford University Press, pp. 256–281.
- [Son*20] Xinfang Song, Wei Jiang, Xiaohui Liu, Hui Lu, Zhihong Tian, and Xiaojiang Du. "A survey of game theory as applied to social networks." In: *Tsinghua Science and Technology* 25.6 (2020). Publisher: TUP, pp. 734–742.
- [Sta85] Ervin Staub. "The psychology of perpetrators and bystanders." In: *Political Psychology* (1985). Publisher: JSTOR, pp. 61–85.
- [Ste*21] Harald Steck, Linas Baltrunas, Ehtsham Elahi, Dawen Liang, Yves Raimond, and Justin Basilico. "Deep learning for recommender systems: A Netflix case study." In: *AI Magazine* 42.3 (2021), pp. 7–18.
- [SK09] Xiaoyuan Su and Taghi M. Khoshgoftaar. "A Survey of Collaborative Filtering Techniques." In: *Advances in Artificial Intelligence 2009* (October 2009). Ed. by Jun Hong. Publisher: Hindawi Publishing Corporation, p. 421425. ISSN: 1687-7470. DOI: 10.1155/2009/421425.
- [Suk*15] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, and others. "End-to-end memory networks." In: *Advances in neural information processing systems* 28 (2015).
- [STY17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." In: *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.
- [SV08] Cass R Sunstein and Adrian Vermeule. "Conspiracy theories." In: (2008). Publisher: Harvard public law working paper.
- [Tac19] Veronika Tacke. "Systeme und Netzwerke-oder: Was man an sozialen Netzwerken zu sehen bekommt, wenn man sie systemtheoretisch beschreibt." In: *Netzwerke und Soziale Arbeit. Theorien, Methoden, Anwendungen*. 2019.
- [Taj70] Henri Tajfel. "Experiments in intergroup discrimination." In: *Scientific american* 223.5 (1970). Publisher: JSTOR, pp. 96–103.
- [Taj74] Henri Tajfel. "Social identity and intergroup behaviour." In: *Social science information* 13.2 (1974). Publisher: Sage Publications Sage CA: Thousand Oaks, CA, pp. 65–93.
- [Taj*79] Henri Tajfel, John C Turner, William G Austin, and Stephen Worchel. "An integrative theory of intergroup conflict." In: *Organizational identity: A reader* 56.65 (1979), pp. 9780203505984–16.
- [Tep12] Peter Tepe. "Ideologie." In: *Ideologie*. de Gruyter, 2012.
- [TBB21] Ludovic Terren and Rosa Borge-Bravo. "Echo chambers on social media: A systematic review of the literature." In: *Review of Communication Research* 9 (2021), pp. 99–118.
- [TM07] Nava Tintarev and Judith Masthoff. "A survey of explanations in recommender systems." In: *2007 IEEE 23rd international conference on data engineering workshop*. IEEE, 2007, pp. 801–810.

- [TM22] Antonela Tommasel and Filippo Menczer. “Do Recommender Systems Make Social Media More Susceptible to Misinformation Spreaders?” In: *Proceedings of the 16th ACM Conference on Recommender Systems*. 2022, pp. 550–555.
- [TS72] Ernst Topitsch and Kurt Salamun. “Ideologie: Herrschaft des Vor-Urteils.” In: *Langen-Müller-Stichworte 5* (1972). Publisher: Langen-Müller.
- [TWVE19] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. “From Louvain to Leiden: guaranteeing well-connected communities.” In: *Scientific reports* 9.1 (2019). Publisher: Nature Publishing Group, pp. 1–12.
- [TB22] Lloyd N Trefethen and David Bau. *Numerical linear algebra*. Vol. 181. Siam, 2022.
- [TSMST08] Karen HL Tso-Sutter, Leandro Balby Marinho, and Lars Schmidt-Thieme. “Tag-aware recommender systems by fusion of collaborative filtering algorithms.” In: *Proceedings of the 2008 ACM symposium on Applied computing*. 2008, pp. 1995–1999.
- [TPH19] Hessel Tuinhof, Clemens Pirker, and Markus Haltmeier. “Image-based fashion product recommendation with deep learning.” In: *Machine Learning, Optimization, and Data Science: 4th International Conference, LOD 2018, Volterra, Italy, September 13-16, 2018, Revised Selected Papers 4*. Springer, 2019, pp. 472–481.
- [Twi23] Twitter. *Twitter’s Recommendation Algorithm*. April 2023. URL: https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm (visited on October 24, 2023).
- [Tö18] Petter Törnberg. “Echo chambers and viral misinformation: Modeling fake news as complex contagion.” In: *PLoS ONE* (2018).
- [Tö22] Petter Törnberg. “How digital media drive affective polarization through partisan sorting.” In: *Proceedings of the National Academy of Sciences* 119.42 (2022). Publisher: National Acad Sciences, e2207159119.
- [UB12] Jodie B Ullman and Peter M Bentler. “Structural equation modeling.” In: *Handbook of Psychology, Second Edition 2* (2012). Publisher: Wiley Online Library.
- [Ulr94] Gunter Ulrich. *Politische Steuerung*. Springer, 1994.
- [VBPC08] Jay J Van Bavel, Dominic J Packer, and William A Cunningham. “The neural substrates of in-group bias: A functional magnetic resonance imaging investigation.” In: *Psychological science* 19.11 (2008). Publisher: SAGE Publications Sage CA: Los Angeles, CA, pp. 1131–1139.
- [Var20] Lav R Varshney. “Respect for human autonomy in recommender systems.” In: *arXiv preprint arXiv:2009.02603* (2020).
- [VLP19] Vítor V Vasconcelos, Simon A Levin, and Flávio L Pinheiro. “Consensus and polarization in competing complex contagion processes.” In: *Journal of the Royal Society Interface* 16.155 (2019). Publisher: The Royal Society, p. 20190196.
- [Vas*17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, \Lukasz Kaiser, and Illia Polosukhin. “Attention is all you need.” In: *Advances in neural information processing systems* 30 (2017).
- [WRM08] Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation schemes*. Cambridge University Press, 2008.

- [WB11] Chong Wang and David M Blei. “Collaborative topic modeling for recommending scientific articles.” In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 448–456.
- [Wan*17] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. “Knowledge graph embedding: A survey of approaches and applications.” In: *IEEE Transactions on Knowledge and Data Engineering* 29.12 (2017). Publisher: IEEE, pp. 2724–2743.
- [Wat04] Duncan J Watts. *Small worlds: the dynamics of networks between order and randomness*. Vol. 36. Princeton university press, 2004.
- [Wei17] Jian Wei. “Collaborative filtering and deep learning based recommendation system for cold start items.” In: *Expert Syst. Appl.* (2017).
- [Wei95] Karl E Weick. *Sensemaking in organizations*. Vol. 3. Sage, 1995.
- [Wei*02] Gérard Weisbuch, Guillaume Deffuant, Frédéric Amblard, and Jean-Pierre Nadal. “Meet, discuss, and segregate!” In: *Complexity* 7.3 (2002). Publisher: Wiley Online Library, pp. 55–63.
- [WCB15] Jason Weston, Sumit Chopra, and Antoine Bordes. “Memory networks.” English (US). In: 2015.
- [Wu*22] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. “Graph neural networks in recommender systems: a survey.” In: *ACM Computing Surveys* 55.5 (2022). Publisher: ACM New York, NY, pp. 1–37.
- [Xia*21] Lianghao Xia, Chao Huang, Yong Xu, Peng Dai, Xiyue Zhang, Hongsheng Yang, Jian Pei, and Liefeng Bo. “Knowledge-enhanced hierarchical graph transformer network for multi-behavior recommendation.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. Issue: 5. 2021, pp. 4486–4493.
- [XSA20] Huihui Xu, Jaromír Savelka, and Kevin D. Ashley. “Using Argument Mining for Legal Text Summarization.” In: *International Conference on Legal Knowledge and Information Systems*. 2020. URL: <https://api.semanticscholar.org/CorpusID:229370937>.
- [Yua*22] Jiangjun Yuan, Jiawen Shi, Jie Wang, and Weinan Liu. “Modelling network public opinion polarization based on SIR model considering dynamic network structure.” In: *Alexandria Engineering Journal* 61.6 (2022). Publisher: Elsevier, pp. 4557–4571.
- [ZLJ20] Guijuan Zhang, Yang Liu, and Xiaoning Jin. “A survey of autoencoder-based recommender systems.” In: *Frontiers of Computer Science* 14 (2020). Publisher: Springer, pp. 430–450.
- [Zha18] Jingjing Zhang. “Exploring Explanation Effects on Consumers’ Trust in Online Recommender Agents.” In: *International Journal of Human–Computer Interaction* (2018).
- [Zha*06] Sheng Zhang, Weihong Wang, James Ford, and Fillia Makedon. “Learning from incomplete ratings using non-negative matrix factorization.” In: *Proceedings of the 2006 SIAM international conference on data mining*. SIAM, 2006, pp. 549–553.
- [Zha*19] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. “Deep learning based recommender system: A survey and new perspectives.” In: *ACM computing surveys (CSUR)* 52.1 (2019). Publisher: ACM New York, NY, USA, pp. 1–38.
- [Zha*20] Yizhou Zhang, Yibao Wang, Tinggui Chen, and Jiawen Shi. “Agent-based modeling approach for group polarization behavior considering conformity and network re-

- relationship strength.” In: *Concurrency and Computation: Practice and Experience* 32.14 (2020). Publisher: Wiley Online Library, e5707.
- [ZC*20] Yongfeng Zhang, Xu Chen, et al. “Explainable recommendation: A survey and new perspectives.” In: *Foundations and Trends® in Information Retrieval* 14.1 (2020), pp. 1–101.
- [ZNY17] Lei Zheng, Vahid Noroozi, and Philip S Yu. “Joint deep modeling of users and items using reviews for recommendation.” In: *Proceedings of the tenth ACM international conference on web search and data mining*. 2017, pp. 425–434.
- [ZZ10] Feng Zhu and Xiaoquan Zhang. “Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics.” In: *Journal of marketing* 74.2 (2010). Publisher: SAGE Publications Sage CA: Los Angeles, CA, pp. 133–148.
- [ZR21] Reza Jafari Ziarani and Reza Ravanmehr. “Serendipity in recommender systems: a systematic literature review.” In: *Journal of Computer Science and Technology* 36 (2021). Publisher: Springer, pp. 375–396.
- [Zub19] Shoshana Zuboff. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, 2019. ISBN: 978-1-61039-569-4.

List of figures

2.1	Screenshot of the prototype RS for the first user study: The current user has weighted the tags “Action” and “Sci-Fi” (a), therefore receiving matching movie recommendations such as “Matrix” or “Star Wars” (d). The user can also search for other tags provided by the users or get inspiration from the suggestions (b). Furthermore, the user’s existing profile is explained by a tag cloud (c).	12
2.2	Screenshot of the prototype RS for the second user study: A user whose profile is shown in the dialog (c) has applied a critique (a) to the currently recommended movie “Apocalypse Now” using the tags “dark” and “comedy”. As a consequence, recommendations that fit to the critique and to his or her long-term interests are shown (d). To add further critique dimensions, the user can also search for tags provided by other users (b).	13
2.3	Normalized movie positions with respect to tag relevance (left) and latent factors (right).	16
2.4	Simplified schematic illustration of the proposed rating prediction pipeline including read operations for the neural memory components.	19
2.5	Exemplary user profile that depicting (assumed) personal and average importance for three aspects as well as the target user’s comments sorted by aspects.	21
2.6	Recommendation for the movie Parasite including the predicted rating, an overview of (assumed) aspect importance, and a personally selected comment that supports the predicted rating. Selection of comments can be personalized by toggling the respective radio button.	22
3.1	Exemplary visualizations of graph topologies and latent space polarization for the partisan scenario. The graphs shown for each condition (figs. 3.1b to 3.1d) are taken from one randomly selected simulation run at final epoch 100. While in the initial graph (fig. 3.1a), the two communities are strongly intertwined, all conditions are able to clearly separate them, with $\mathcal{L}_2 \times \text{Personalized}$ achieving the strongest delineation. Condition $\mathcal{L}_1 \times \text{Random}$ is omitted because the graphs shows no notable effects. Below (figs. 3.1e to 3.1h), polarization in latent space is depicted in the form of joyplots (also known as ridgeline plots) giving a sense of the evolution of polarization over time (y-axis). The plots show the kernel density estimations of a one-dimensional principal component analysis (PCA) on the latent representations of user nodes. The color gradient from yellow (low) to violet (high) depicts the PCA’s explained variance. The results reveal that the slight separation at the start cannot be maintained by conditions $\mathcal{L}_1 \times \text{Random}$ and $\mathcal{L}_1 \times \text{Personalized}$, while the remainder increase bipolarization, with $\mathcal{L}_2 \times \text{Personalized}$ showing the most pronounced effects.	61

- 3.2 Exemplary visualizations of graph topologies and latent space polarization for the divisive scenario. The graphs shown (figs. 3.2b to 3.2d) are each taken from one simulation run at epoch 50. Initially (fig. 3.2a), two small communities (depicted in violet and yellow) are integrated to some degree into a larger community (depicted in grey) while forming at opposite edges of the social graph. In condition $\mathcal{L}_1 \times$ Personalized, both peripheral communities spread out, while in $\mathcal{L}_2 \times$ Random, they become further integrated into the grey community, thus loosening internal connectivity. Concerning condition $\mathcal{L}_2 \times$ Personalized, we observe pronounced centrifugal tendencies with condensed structures forming separately in the focal points of the respective communities. Below (figs. 3.2e to 3.2h), polarization in latent space is depicted. Due to the presence of not only two, but three communities, we depict latent user distributions in form of a three-dimensional scatter plot. Concretely, we show the migration of user distributions by plotting a two-dimensional PCA over time at epochs 1, 25, 50, and 100. Clearly, only condition $\mathcal{L}_2 \times$ Personalized shows bipolarization while the remaining converge to consensus. 63
- 3.3 Visualizations of graph topologies on a real-world dataset. The initial graph (fig. 3.3a) consists of 20% of the earliest edges contained in the target graph (fig. 3.3b). While the initial graph has very loose connections and thus a large number of communities is identified, the target graph is much more condensed and contains only 5 communities. We calculated Jaccard similarity to the target network's communities to assign colors to the communities in the initial as well as the scenario graphs. All three depicted scenarios (figs. 3.3c to 3.3e) capture to some degree the properties of the target graph. For instance, they all identify the most clearly separated communities (in orange and yellow). However, at closer inspection $\mathcal{L}_1 \times$ Personalized shows too much and $\mathcal{L}_2 \times$ Random too little condensation. Although there are still obvious topological discrepancies to the target graph, we view $\mathcal{L}_{1|2} \times$ Personalized as achieving the closest approximation. 66

List of tables

- 2.1 Example for automatically learned relationships between latent factors (rows) and user-generated tags (columns): The five most important factors are shown together with positively (yellow) and negatively (purple) related tags. The factor importance is depicted in brackets in the left-most column. Representatives for each factor are automatically determined by extracting the movies (with at least 10 000 ratings) that score highest for the respective factor in the actual item-factor matrix. 15

The following article is reused from:

Donkers, T., Loepp, B., & Ziegler, J. (2017). Sequential User-Based Recurrent Neural Network Recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (pp. 152-160). Association for Computing Machinery.
<https://doi.org/10.1145/3109859.3109877>

Sequential User-based Recurrent Neural Network Recommendations

Tim Donkers
University of Duisburg-Essen
Duisburg, Germany
tim.donkers@uni-due.de

Benedikt Loepp
University of Duisburg-Essen
Duisburg, Germany
benedikt.loeppl@uni-due.de

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

Recurrent Neural Networks are powerful tools for modeling sequences. They are flexibly extensible and can incorporate various kinds of information including temporal order. These properties make them well suited for generating sequential recommendations. In this paper, we extend Recurrent Neural Networks by considering unique characteristics of the Recommender Systems domain. One of these characteristics is the explicit notion of the user recommendations are specifically generated for. We show how individual users can be represented in addition to sequences of consumed items in a new type of Gated Recurrent Unit to effectively produce personalized next item recommendations. Offline experiments on two real-world datasets indicate that our extensions clearly improve objective performance when compared to state-of-the-art recommender algorithms and to a conventional Recurrent Neural Network.

CCS CONCEPTS

• Information systems → Recommender systems; • Computing methodologies → Artificial intelligence;

KEYWORDS

Recommender Systems, Deep Learning, Neural Networks, Recurrent Neural Networks, Sequential Recommendations

ACM Reference format:

Tim Donkers, Benedikt Loepp, and Jürgen Ziegler. 2017. Sequential User-based Recurrent Neural Network Recommendations. In *Proceedings of RecSys '17, Como, Italy, August 27-31, 2017*, 9 pages. <https://doi.org/10.1145/3109859.3109877>

1 INTRODUCTION

The majority of today's *Recommender Systems* (RS) [41] relies on algorithms that are designed under the assumption

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

RecSys '17, August 27-31, 2017, Como, Italy

© 2017 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.

ACM ISBN 978-1-4503-4652-8/17/08...\$15.00

<https://doi.org/10.1145/3109859.3109877>

that user preferences are static patterns [29, 50]. *Collaborative Filtering* (CF) [28] approaches, both neighborhood- and model-based, have proved immensely useful without any consideration of time. However, assuming consumption events to be independent from each other precludes taking advantage of the temporal dynamics that naturally exist in user behavior [50]. Thus, despite the practical success of time-agnostic models, it seems promising to investigate whether the inclusion of temporal aspects can improve effectiveness. For instance, many commercial goods are only bought during a specific season. Music songs are played in succession according to, among others, the user's current mood or the desire for diversity, and are often arranged in playlists. Fictional series are typically consumed one episode after another, before any other content is considered. In contrast to the conception prevailing in many RS, individual data points used in temporal or sequential recommendations thus cannot be assumed independent and uniformly distributed, as they typically form correlated sequences. Several ways exist to approach this issue, e.g. by extending conventional time-independent algorithms [e.g. 7, 27], integrating time as a contextual factor [e.g. 22, 24], or by applying sequence-based methods primarily used in other domains [e.g. 42]. Still, many approaches ignore that generating recommendations is an inherently time-dependent task and focus only on achieving increasingly accurate predictions.

Compared to other areas, sequence modeling has overall been less explored in RS research due to the long-lasting focus on time-agnostic models. This gap may also be explained by the unique nature of the recommendation problem. Although heavily influenced by Machine Learning [41], many modern techniques successfully applied in other domains cannot easily be transferred to RS. This is particularly true for techniques from a research field currently attracting much attention, *Deep Learning* (DL). DL allows to solve problems of representation learning, i.e. automatically discovering adequate representations of data without manual feature engineering [10]. These representations are expressed in terms of stacked, hierarchically ordered simpler representations. While DL is widely used with considerable success in areas such as Natural Language Processing [6, 11, 12, 33, 36] or Image Processing [16, 30, 47], it has only rarely been applied to the recommendation problem. The few exceptions focus on purposes other than sequential recommending, e.g. using DL as a preprocessing step to conventional techniques [49, 52]. Even if they consider consumption sequences [13, 18, 19, 48, 53, 54], they often have other limitations.

Recurrent Neural Networks (RNN) represent a specific form of DL models which possess several properties that make them attractive for sequence modeling [10]. In particular, they are capable of incorporating input from past consumption events, allowing to derive a wide range of sequence-to-sequence mappings. In order to consider time as a first class factor for modeling user preferences in RS, RNNs thus constitute a promising family of techniques. One shortcoming of simple variants, however, is the limited number of input variables that can effectively be handled [2]. Gated architectures [6] are designed to overcome this limitation by including gating units trained to control information flow through the network, thereby learning to keep information over a long period of time. Both *Long Short-Term Memory* (LSTM) [9, 20] and *Gated Recurrent Unit* (GRU) [5, 6] networks have shown advantages in real-world applications. Gated RNNs, however, have not been designed with the recommendation domain in mind. In particular, they are not optimized for taking interaction between user and system into account. As mentioned above, first attempts have been made to use DL for sequential recommending [13, 18, 19, 48, 53, 54]. Yet, they tend to reuse standard networks without modifications. As a result, the models usually have no notion about the specification of a target sequence, i.e. that respective RS have no mechanism to consider the fact that a sequence of consumed items is unique to an individual user. The networks treat every single sequence equivalently, thus learning global consumption behavior, but are unable to (deeply) integrate user-specific information.

In this paper, we propose a novel DL approach to address the sequential recommendation problem. First, relying on the benefits of gated RNNs, we model the temporal dynamics of consumption sequences. Second, through a novel gated architecture with additional input layers, we explicitly represent the individual user in such a network. These user-based GRUs are uniquely designed and optimized for the purpose of generating personalized next item recommendations. We theoretically and empirically show that not only out-of-the-box RNNs, but especially our approach using a novel gated architecture, may outperform state-of-the-art recommender algorithms as well as their time-dependent counterparts with respect to objective performance when predicting sequences. Offline experiments conducted on two real-world datasets indicate significant improvements when comparing our approach to common baseline RS algorithms and to a conventional RNN.

The remainder of this paper is organized as follows: First, we discuss relevant work related to sequential recommending and to applying DL in RS research. Second, we present our approach of user-based RNNs for RS. Next, we describe how we evaluated this approach in offline experiments. Finally, we conclude the paper and discuss potential future work.

2 RELATED WORK

User representation in RS is often designed as a statistically stationary process where every expressed preference is assumed to be fixed over time [29, 50]. However, relational data in real-world scenarios are often evolving and exhibit strong temporal patterns [55]. Time may therefore be considered an important contextual dimension also in RS [4]. Generating recommendations in a time-independent manner may in contrast result in estimations of preferences based on user-item relations that are no longer valid. This lack of adaptability with respect to one of the most natural properties of user behavior has motivated several approaches that integrate temporal dynamics¹. In the following, we discuss the most important ones, and especially those most closely related to ours that also exploit DL techniques as a means to generate sequential recommendations.

2.1 Towards Sequential Recommendations

To adequately reflect user interests, functions that rank items according to some notion of utility are today considered most suitable for generating personalized recommendations [14]. Advances in RS research suggest that implicit feedback thereby often provides more comprehensive and extensive insights into user behavior than explicit ratings [39]. Fitting a static model to data naturally depending on dynamic processes approximates an improper data-generating function [42]. For this reason, several methods exist to incorporate temporal information into RS at some stage, which can be differentiated as follows [50]:

- *Time-aware RS* consider time as a contextual feature during the training phase. Timestamps serve as an additional source of information by which the model is enriched. The rationale behind is that user behavior underlies certain habits and regularities that repeat in regular time intervals, consequently allowing a more accurate prediction of similar patterns in the future.
- *Time-dependent RS* consider user preference data as chronologically ordered sequences, assuming that the most intrinsic property is that time establishes an order of the events. Input is required in chronological form, while exact time spans do not need to be taken into account. The algorithms consequently do not aim at modeling time as being cyclic, but rather at adapting to changes.

Existing time-aware methods include pre- and post-filtering [1] as well as Tensor Factorization [24]. In general, statistical models explicitly designed to predict sequences as used, for example, in automated playlist generation [3], may also be applied. However, they typically do not aim at modeling higher-order long-term dependencies. Eventually, time-dependent RS are conceptually closest to the definition of time used throughout this paper. Therefore, in the following, we focus on attempts that utilize chronological input data. By incrementally training models based on recency, the respective

¹See one of the extensive surveys for an overview [e.g. 4, 50].

systems may capture, among others, continuous temporal dynamics such as concept drifts [26], changes in item popularity, or virtually any effect observable in sequential data.

Many approaches extend conventional CF techniques, for instance, by translating k -NN methods. In [18], it has been parenthetically shown that an unpersonalized variant of item-based k -NN achieves reasonably good results in sequence prediction tasks. Under an assumption similar to the Markov property, each prediction only depends on the item the user has previously consumed. In [7], a fully personalized time-discounted version has been proposed by extending k -NN with an exponential decaying component that penalizes items consumed a long time ago. Other approaches use time as a means for determining similarities, e.g. when consumption events are close together [17] or lie within a sliding window employed over the most recent sessions [35].

Matrix Factorization (MF) [29] algorithms have also been adapted. For example, baseline predictors can be added to the original framework to account for temporal dynamics [27]. Similar to their time-aware counterparts, sequential MF models have been extended to Tensor Factorization [55]. The dimensions contain time intervals so that the tensor can be seen as a periodic collection of ratings over time. Others have turned the MF problem into a graph-based model [38], integrated a Markov model for predicting the next shopping basket [40], or generalized existing temporal MF approaches to latent time series models [57]. Overall, the variant proposed in [23] comes closest to the definition we use: The authors model temporal order by means of a time window that includes feature vectors of previously consumed items. Thus, both collaborative interactions and time series aspects of the collaborative data can be taken into account.

2.2 Deep Learning in RS Research

The approaches discussed before constitute extensions of established time-agnostic recommendation methods. DL techniques such as RNNs can also be considered useful for sequential recommending as they are specifically designed for modeling temporal aspects. Despite the advantages of such techniques, only few neural network approaches have yet been generally proposed in RS research—most of them not focused on sequential recommendations.

One of the first attempts that only uses a shallow network structure is presented in [43]. Here, two-layered undirected graphical models called Restricted Boltzmann Machines are used to model user ratings for the prediction task. Deep neural networks have in contrast primarily been applied in preprocessing steps to conventional RS techniques. In [49], for instance, features are extracted from unstructured content, here raw music data. These features then serve as input to conventional CF in order to improve its performance. In [52], DL is performed on generic content-based information. The resulting deep feature representation is used as an enhancement to address the sparsity problem occurring in CF. More specific data, e.g. tag-based user and item profiles, have been embedded into a deep feature space in order to approach the

hardly controllable dynamics of underlying tag corpora [56]. For that, the authors model similarities between users and potential target items by inferring a deep semantic model. Finally, some approaches actually exploit user interaction behavior. For example, for the purpose of generating adaptive user interfaces, i.e. recommending next actions based on similar interaction patterns found in the user base, GRUs have been used to embed learned interaction patterns together with users and interaction elements [46].

Only few works have yet used DL to explicitly generate sequential recommendations in a self-contained manner. The proposed approaches thereby often only focus on modeling consumption sequences without explicitly representing individual users [13, 18, 19, 48]. For instance, the applicability of the approach presented in [18] is restricted to sessions only. While user representation remains uncovered, the authors introduce a variant of a GRU model that utilizes pair-wise loss functions. In [48], this approach is extended by means of a priori data augmentation in form of sequence preprocessing. In addition, to reduce influence of outdated properties, essentially two models are trained: The first one on the complete dataset, which is then used to initialize the second one subsequently trained only on a subset of newer samples. Only recently, the RNN from [18] has been merged with feature-rich content information [19]: Item one-hot vectors and vectors of extracted features (e.g. corresponding to images of the respective items) are simultaneously treated as input to a GRU layer.

Few exceptions also consider user-related aspects [53, 54], but lack a deep integration of user vectors into the gating process. In [53], the RNN framework is extended by solving two recommendation problems and then merging the outputs: First, information is recurrently extracted from consumed item sequences. Second, specific user concepts are distinguished in a feed-forward manner. As a result, user characteristics are only considered independently from sequence properties. In [54], the authors train individual RNNs to model user and item evolution separately. The outputs from both networks are subsequently coupled with further auxiliary parameters capturing stationary concepts in order to predict user ratings. This architecture requires learning two RNNs such that user and item properties can yet again only loosely be intertwined.

3 USER-BASED RECURRENT NEURAL NETWORKS

Generating sequential recommendations relies on the assumption that individual data points form correlated sequences. Vectorized abstractions of user preferences or behavioral patterns are valuable information sources that can help to improve recommendation quality. In the previous section, we have discussed several existing methods for formalizing a sequential problem in the context of time-dependent recommendations. In the following, we elaborate on how we consider temporal dynamics in RS using DL techniques, and present a novel gated architecture for RNNs using specialized

GRUs that allows us to seamlessly integrate user-related information into the model.

The explicit notion of user information distinguishes our work from other DL approaches proposed for RS that usually focus on consumption sessions without explicitly representing the user. Compared to the few exceptions that already integrate user-related information (see Section 2), our approach is the first to deeply integrate user vectors into the gating process.

3.1 Recurrent Neural Networks for RS

First, for generating sequential recommendations, we need to concretize the generalized, domain-independent formulation of RNNs and transfer it to RS. RNNs, especially when augmented with gating layers, are powerful tools for modeling sequences of any kind². We rely on a variant of RNNs that produces an output $\mathbf{o}^{(t)} \in \mathbb{R}^{|I|}$ at each time step via an affine transformation of the current hidden state $\mathbf{h}^{(t)} \in \mathbb{R}^n$:

$$\mathbf{o}^{(t)} = T_{n,|I|} \mathbf{h}^{(t)}, \quad (1)$$

where $T_{k,l} : \mathbb{R}^k \rightarrow \mathbb{R}^l$ is an affine transformation of the form $\mathbf{W}\mathbf{x} + \mathbf{b}$. In conventional RNNs, the computation of a hidden state is defined as a function of the previous hidden state and an input vector. In the context of RNNs for RS, we define this input vector $\mathbf{i}^{(t)} \in \{0, 1\}^{|I|}$ as a one-hot vector where the only index different from zero corresponds to the index of a particular item.

3.2 User-based Gated Recurrent Units

Now, in order to integrate user characteristics into the model, we first define a one-hot user variable $\mathbf{v}^{(t)} \in \{0, 1\}^{|Y|}$ with Y being the set of users. We assume the data to be comprised of user-item tuples such that \mathbf{v} can be interpreted as an indicator of the user consuming item \mathbf{i} at a certain time step t in the sequence. Based on this, we extend the original definition (see [e.g. 6]) of the hidden state:

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{i}^{(t)}, \mathbf{v}^{(t)}; \boldsymbol{\theta}) \quad (2)$$

The network's predictive power may benefit from explicitly memorizing concepts about the user inside the recurrent cells. In the following, we discuss several realizations of (2) that architecturally modify the original recurrent unit³. By incorporating a user variable $\mathbf{v}^{(t)}$, we can thus deeply integrate user-related information into the network.

3.2.1 Linear User Integration. Linear user integration is comparable to strategies of considering context for gated units. For instance, in [34], word-related topic vectors are incorporated into a neural network language model to improve performance. Similarly, $\mathbf{v}^{(t)}$ can be viewed as an additional input layer that is connected to the gated units. Thereby, it can influence decisions about forgetting certain pieces of

information or updating hidden states:

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{r} \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \end{bmatrix} T_{3n,2n} \begin{bmatrix} \mathbf{h}^{(t-1)} \\ \mathbf{E}_i \mathbf{i}^{(t)} \\ \mathbf{E}_v \mathbf{v}^{(t)} \end{bmatrix}, \quad (3)$$

where σ is the logistic sigmoid. $\mathbf{E}_i \in \mathbb{R}^{|I| \times n}$ and $\mathbf{E}_v \in \mathbb{R}^{|Y| \times n}$ are embedding matrices that map one-hot vectors into a densified information space. This kind of densification via projection is widely applied in DL research (see [e.g. 32, 33]).

While the gating processes inside the hidden units deal with temporal aspects of the data, non-temporal structures can be exploited by including embedding. In the embedded space, items that appear in similar contexts are also spatially close. For notational simplicity, we assume that the embedding dimension is equal to the hidden dimension n although this does not necessarily has to be the case.

Besides influencing update and reset gates, the user vector can also be included in the calculation of hidden states. For this, we first calculate a state update vector $\mathbf{k} \in \mathbb{R}^n$:

$$\begin{aligned} \mathbf{k} = & \tanh(T_{n,n} \mathbf{r} \odot \mathbf{h}^{(t-1)} \\ & + T_{n,n} \mathbf{E}_i \mathbf{i}^{(t)} \\ & + T_{n,n} \mathbf{E}_v \mathbf{v}^{(t)}), \end{aligned} \quad (4)$$

where \tanh is the hyperbolic tangent and \odot represents element-wise multiplication.

Subsequently, we can leakily integrate \mathbf{k} into the hidden state update mechanism subject to \mathbf{u} :

$$\mathbf{h}^{(t)} = (\mathbf{1}_n - \mathbf{u}) \odot \mathbf{h}^{(t-1)} + \mathbf{u} \odot \mathbf{k}, \quad (5)$$

where $\mathbf{1}_n$ is a vector of ones.

While $\mathbf{i}^{(t)}$ changes with every time step, consumption sequences are unique to only a single user, i.e. $\mathbf{v}^{(t)}$ remains constant. Generalized properties of each item in the sequence relate to the consuming user by feeding back succeeding losses to parameters conditioned by this particular $\mathbf{v}^{(t)}$. As a result, trained user-related parameters are representing superordinate concepts. This means, parameters corresponding to $\mathbf{v}^{(t)}$ express the user's general preference structure. Figure 1 depicts such a linear user GRU cell.

3.2.2 Rectified Linear User Integration. The linear integration of $\mathbf{v}^{(t)}$ allows the recurrent cell to condition the activation based on user characteristics. However, repeatedly applying the same user vector increases redundancy that might lead to undesired effects like higher sensitivity to underfitting and non-zero predictions for all user-item pairs since the particular network parameters are trained with respect to unchanging input over a long series of steps.

Furthermore, user-related parameters might not be equally important at every step in time. Some recommendations might be sufficiently derived only based on item vectors without considering a user component. For example, one can assume that users will consume all parts of a movie trilogy in a row. Thus, cells should learn to selectively set focus on different parts of the user representation. We use a leaky

²For a general introduction to RNNs, please refer to [e.g. 31].

³We use GRUs due to notational simplicity, although all principles can easily be transferred to LSTM in analogous form.

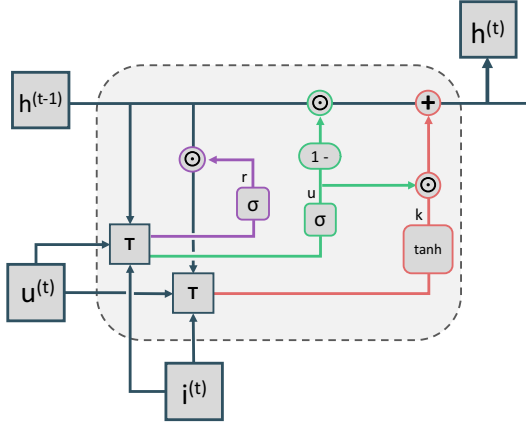


Figure 1: Linear user-based GRU cell: In addition to the item input vector \mathbf{i} , the user representation \mathbf{v} is included.

integrator inspired by rectified linear units [15] to attach weight to the transformed components of $\mathbf{v}^{(t)}$:

$$(\mathbf{E}_v \mathbf{v}^{(t)})_l \leftarrow \begin{cases} 0, & \text{if } (\mathbf{E}_v \mathbf{v}^{(t)})_l < \kappa_{1,l} \\ \omega (\mathbf{E}_v \mathbf{v}^{(t)})_l, & \text{if } \kappa_{1,l} < (\mathbf{E}_v \mathbf{v}^{(t)})_l < \kappa_{2,l} \\ (\mathbf{E}_v \mathbf{v}^{(t)})_l, & \text{else} \end{cases} \quad (6)$$

where l refers to a particular entry of vector $\mathbf{E}_v \mathbf{v}^{(t)}$. $\kappa_1 \in \mathbb{R}^n$ and $\kappa_2 \in \mathbb{R}^n$ are threshold parameter vectors conditioned on previous hidden state as well as item and user vector:

$$\begin{bmatrix} \kappa_1 \\ \kappa_2 \end{bmatrix} = T_{3n,2n} \begin{bmatrix} \mathbf{h}^{(t-1)} \\ \mathbf{E}_i \mathbf{i}^{(t)} \\ \mathbf{E}_v \mathbf{v}^{(t)} \end{bmatrix} \quad (7)$$

The thresholds dynamically ensure that only relevant parts of the user representation are exploited to produce network output. User-specific concepts can thus be shut off if there is reason to assume that the current input should be handled independently. Note that very important concepts, i.e. high values in $\mathbf{E}_v \mathbf{v}^{(t)}$, are only discarded if there is a strong indication of orthogonality, i.e. high values for κ_1 .

In cases where a component is only important to a certain degree, i.e. $\kappa_{1,l} < (\mathbf{E}_v \mathbf{v}^{(t)})_l < \kappa_{2,l}$, it might be advantageous not to turn it off completely, but only to limit its throughput. For this purpose, we also integrate a leaky variant with $\omega \in [0, 1]$. Update and reset cells are then calculated according to (3) with the filtered user component. The leaky integrator is included in the cell shown in Figure 2.

3.2.3 Attentional User Integration. As already derived in Section 3.2.1, the user component represents stable concepts as opposed to the current input vector $\mathbf{i}^{(t)}$ that rather expresses short-term aspects. The rectified linear user integration is only designed to impede the user component when necessary. As an alternative, the network could adaptively shift focus between user and item aspects. For instance, the

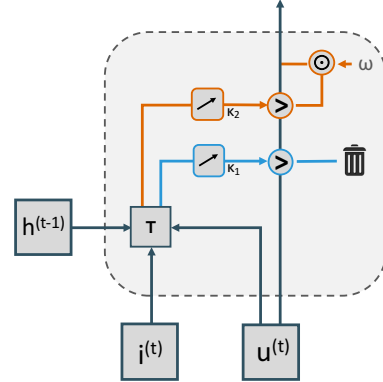


Figure 2: The rectified linear gating process detached from the complete cell: Item and user vectors as well as previous hidden state are concatenated and mapped linearly twice in order to calculate κ_1 and κ_2 . Components of the user vector that are element-wise smaller than the corresponding values in κ_1 are discarded. Components smaller than κ_2 are diminished by element-wise multiplication with ω .

user component's influence should be low at $t=0$ when no information about the user is present, and progressively increase afterwards as the sequence is further propagated. Moreover, at times, a consumed item might be out of the ordinary and would decrease succeeding estimation quality. In such cases, resorting to a stable user representation and excluding the outlier could attenuate or even annul its impact.

We therefore propose an adaptive approach that includes a new kind of gated cell that regulates the gating process. Let $\xi \in \mathbb{R}^n$ be the attentional regulation gate:

$$\xi = \sigma(T_{n,n} \mathbf{h}^{(t-1)} + T_{n,n} \mathbf{E}_i \mathbf{i}^{(t)} + T_{n,n} \mathbf{E}_v \mathbf{v}^{(t)}) \quad (8)$$

We use logistic sigmoid as a squashing function to leakily regulate the proportion of item and user focus. In contrast to a linear user-based GRU cell, item and user vectors are now weighted by ξ :

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{r} \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \end{bmatrix} T_{3n,2n} \begin{bmatrix} \mathbf{h}^{(t-1)} \\ (1_n - \xi) \odot \mathbf{E}_i \mathbf{i}^{(t)} \\ \xi \odot \mathbf{E}_v \mathbf{v}^{(t)} \end{bmatrix} \quad (9)$$

Furthermore, the hidden state's update mechanism is not only dependent on $\mathbf{i}^{(t)}$ and $\mathbf{v}^{(t)}$, but also on ξ :

$$\begin{aligned} \mathbf{k} &= \tanh(T_{n,n} \mathbf{r} \odot \mathbf{h}^{(t-1)} \\ &\quad + T_{n,n} (1_n - \xi) \odot \mathbf{E}_i \mathbf{i}^{(t)} \\ &\quad + T_{n,n} \xi \odot \mathbf{E}_v \mathbf{v}^{(t)}) \end{aligned} \quad (10)$$

The state update vector \mathbf{k} is then integrated analogously to (5). In contrast to (4), the attentional gate ξ now acts as a leaky integrator that can choose to completely ignore the user aspect (extremely low sigmoids) or simply copy it (extremely high sigmoids), see Figure 3. Informally speaking, the gate regulates the extent to which users are considered as opposed

to items. Note that parameters for the attention gate should be initialized with values close to zero such that the network behaves similar to a standard GRU at the beginning of the training phase, and then may gradually shift focus. Thus, the network will not explicitly focus on the user component until it has learned to do so.

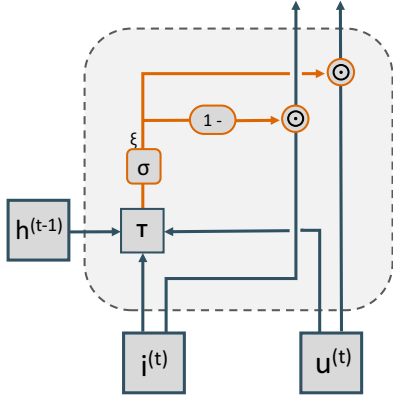


Figure 3: The attentional layer ξ detached from the complete cell: Item and user vectors as well as previous hidden state are concatenated and mapped with sigmoid squashing to form a gate that controls information flow between item and user component. Low sigmoids increase the throughput of item-related components, high sigmoids support the user side.

4 EVALUATION

To objectively justify our theoretical models derived in the previous section, we conducted several offline experiments. In particular, we aimed at answering the question how our novel user-based RNN approach with its different variants to deeply integrate user information performs on the sequential recommendation task when compared to a conventional RNN as well as state-of-the-art recommendation techniques.

Hence, we compared our three user-based networks (i.e. using linear, rectified linear, and attentional user integration) with a standard GRU network without any of our proposed extensions⁴. We also considered the following baselines: We trained time-independent item-based k -NN as well as its time-dependent exponentially decaying counterpart [7]. Moreover, using the BPR learning criterion [39], we trained a standard MF [29] as well as a sequential MF [23] variant.

4.1 Methodology

In the following, we describe the evaluation metrics and datasets we used, as well as the algorithmic setup.

⁴We implemented the different variants using *TensorFlow* (<https://www.tensorflow.org/>), an open source software framework especially designed for building deep neural networks as data flow graphs. For all experiments, we used ECP P2 instances provided by *Amazon Web Services* (<https://aws.amazon.com/>), which are powerful scalable instances designed for GPU-based operations using CUDA.

4.1.1 Metrics. Based on temporally ordered lists of consumed items, our objective is to correctly predict the next item a target user will likely consume. The ground truth at a particular time step is therefore represented by a single user-item tuple. To present the user with adequate recommendations, the target item should be among the first few recommended items. In accordance with recent RS research, we thus use the following evaluation metrics:

- **MRR@ k (Mean Reciprocal Rank)** is defined as the average of the reciprocal ranks of the desired items [51]. The rank is set to zero if it is above k .
- **Recall@ k** is defined as the fraction of cases where the item actually consumed in the next event is among the top k items recommended [37].

We set $k = 20$, as it appears desirable from a user's perspective to expect the target among the first 20 items [18].

4.1.2 Datasets. We ran our approach as well as all the baselines on the following two real-world datasets:

- **MovieLens 10M:** The MovieLens 10M dataset⁵ consists of 10 000 054 ratings assigned to 10 681 movies by 71 567 users. In order to mimic implicit data, we binarized all ratings independent of their value, considering them as positive feedback as it has been done in [e.g. 39]. Using the timestamps provided, we thus got an ordered sequence of consumption events for each user. The dataset contains only sequences with a minimum length of 20, making further preprocessing of the training set unnecessary. The average sequence length is 115. We aimed at predicting the next movie to watch.
- **LastFM 1K Users:** The LastFM 1K Users dataset⁶ contains user-timestamp-artist-song tuples collected via the LastFM API. The dataset has a total of 19 150 868 data points for 992 users. Due to computational reasons we performed our evaluation on a 10% subsample. The resulting average sequence length is 1 738. We aimed at predicting the next song title.

We split each dataset into three parts: First, the major fraction of every sequence serves as training data. Second, we use a validation set during training to measure performance on unseen data. These validation results determine, for instance, early stopping used in the learning phase to avoid overfitting [21]. Third, we use a distinct test set to measure actual performance after learning is completed. The training set consists of the first 90% of every user consumption sequence. The remaining 10% of each sequence are split evenly into validation and test set while maintaining the order. As it is common practice in RS research, items not seen during training are filtered out from validation and test set.

4.1.3 Hyperparameter and Network Setup. Table 1 shows all hyperparameters we set for our experiments based on extensive pretesting with grid search.

For the RNN variants, we propagate mini-batch based learning with batch size of 1000. We use shallow networks

⁵<https://grouplens.org/datasets/movielens/>

⁶<http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/>

Table 1: Hyperparameter values for all algorithms used in our experiments on the two datasets.

Parameter	Value	Parameter	Value
Item-based k-NN		General GRU	
Nearest Neighbors	100	Batch size	1000
Exp. Dec. Item-based k-NN		Layers	1
Decaying Constant	1	Hidden Units	1000
Matrix Factorization		Unfold Dimension	20
Latent Factors	20	Dropout	0.8
Iterations	30	Gradient Cap	5.0
Learning Rate	0.01	Iterations	10
Seq. Matrix Factorization		Learning Rate	0.001
Window Size	2	User-based GRU	
		User Dropout	0.5

of depth 1 with hidden dimensionality 1000 that is unfolded for 20 time steps. Stacking multiple hidden recurrent layers on top of each other to create depth [8, 12, 44] did not improve overall performance in our pretests. According to [36], depth in RNNs introduces capabilities to represent different timescales. When trained on the datasets we used, the networks however did not seem to benefit from this property.

For regularization, we apply Dropout [58] with a keep probability of 0.8 for item and 0.5 for user vectors. We use a gradient cap of 5.0 to clip large gradients that might otherwise lead to skip over minima. Furthermore, we introduce a L^2 parameter norm penalty on user side in order to control for redundancy. Its contribution weight is set to 0.01. Regarding the matrices embedding the one-hot encoded input vectors, we set the dimension equal to the hidden dimension. Weight matrices and biases are initialized randomly with values drawn from $[-0.1, 0.1]$. We run training for 10 iterations.

Linear and attentional user integration do not require setting any further parameters. In case of rectified linear user integration, grid search indicated to set $\omega = 0.2$, i.e. diminished user-related concepts are only considered by 20% of their original magnitude.

Pretests also indicated that the RNNs are trained best with Adam optimizer [25] using a learning rate of 0.001 against Cross Entropy loss. For this, we normalize output vectors $\mathbf{o}^{(t)}$ via softmax function in order to produce a valid probability distribution.

4.2 Results

Table 2 shows the results received after training the baselines as well as our DL models on both datasets.

The RNNs yield results superior to the baselines in terms of both metrics. On the MovieLens dataset, there is a substantial gain in objective performance compared to the next best baseline. Namely, there is an improvement of 12% in MRR@20 and 23% in Recall@20, respectively, when comparing exponentially decaying item-based k -NN to the standard GRU. On the LastFM dataset, the differences in ranking quality become even more apparent (at least 157% improvement in terms of MRR@20, 21% in Recall@20).

Integrating user information significantly improves statistical power in GRU networks. All user-based RNN variants

Table 2: MRR@20 and Recall@20 for all baselines as well as our proposed user-based RNN variants.

	MovieLens		LastFM	
	MRR@20	Recall@20	MRR@20	Recall@20
Baselines				
Item-based k -NN	0.036 39	0.121 42	0.044 38	0.101 26
Exp. Dec. Item-based k -NN	0.042 31	0.128 53	0.055 94	0.167 24
Matrix Factorization	0.011 92	0.077 44	0.032 89	0.127 23
Seq. Matrix Factorization	0.015 50	0.107 30	0.027 69	0.112 24
Standard GRU	0.047 30	0.157 73	0.143 74	0.202 69
User-based RNN				
Linear User-based GRU	0.052 51	0.160 28	0.182 52	0.249 20
Rectified Linear User-based GRU	0.059 01	0.186 97	0.191 54	0.254 09
Attentional User-based GRU	0.062 55	0.205 40	0.187 31	0.254 85

outperform the standard GRU network that does not consider user information. In particular, the integration variant that performs best in terms of MRR@20, i.e. attentional on the MovieLens and rectified linear on the LastFM dataset, leads to an improvement of 32% or 33%, respectively. For Recall@20, there is similar gain with results 30% better on the MovieLens, and 28% better on the LastFM dataset.

Taking a closer look at the results for the user-based RNNs, attentional user integration achieves the overall best results. Only on the LastFM dataset, MRR@20 is slightly lower than for the rectified linear variant. In all other cases, the attentional realization outperforms the others.

Finally, concerning the baselines, item-based k -NN is better at modeling sequences than MF. Even sequential MF achieves clearly inferior results than the original item-based k -NN without any temporal extensions. As a side note, while the temporal item-based k -NN generally outperforms its time-independent counterpart, sequential MF seems not beneficial compared to standard MF on the LastFM dataset.

4.3 Discussion

Overall, the results indicate that RNNs clearly outperform other recommending approaches when it comes to generating sequential recommendations. As earlier experiments by others suggested [e.g. 18, 48], even out-of-the-box RNNs achieve superior results with respect to widely used evaluation metrics when compared to state-of-the-art item-based k -NN or MF approaches. This is particularly true for their derivatives specifically extended to consider temporal effects.

By deeply integrating user information, we were able to personalize the sequence prediction task. The experiments suggest that our networks are able to learn the implicit relations between events more effectively when passing in externally encoded information about users. Nonetheless, the advantages of exploiting temporal dynamics in combination with user information seem to depend on the background data. For instance, the relative performance gain is much higher on the LastFM than on the MovieLens dataset. Considering the nature of the datasets, this is however not surprising: During one session, users generally listen to more than one song in a row, e.g. they consume playlists or whole albums at once, and factors such as the user's mood determine which

title is likely to be consumed next. Thus, similar to the underlying grammar in Natural Language Processing, there exist inherent dependencies between succeeding events that can be exploited. Such relations are much less obvious when recommending movies as people in this case generally do not consume multiple items in immediate succession. Furthermore, sequences in the MovieLens dataset can be considered more artificially, as they do not describe actual consumption events at certain points in time, but reproduce the process of actively providing feedback in form of ratings. Still, there seem to be some implicit connections between the events that can be captured by the networks: Even with such a dataset, RNNs achieve superior results for predicting the next item compared to baseline techniques.

Regarding the different variants of user integration, rectified linear and attentional realization yield generally better results than the linear one. This indicates that regulating redundancy of user input in fact appears to be beneficial. The mechanism seems to support cells in controlling information flow more accurately. In both of the advanced variants, influence of the user variable can be diminished when necessary, thus restricting its impact and preventing overgeneralization. Thereby, it is important to note that integrating user information does not come along with considerably higher runtime requirements. In particular, the user-based RNNs can be trained in almost same amount of time as a standard GRU network. For instance, on the MovieLens dataset the number of processed input tuples per second was on average 24 917 for the Standard GRU, 22 887 for the Linear User-based GRU, 22 840 for the Rectified Linear User-based GRU, and 22 822 for the Attentional User-based GRU. In general, as earlier work on RNNs in RS research has shown [e.g. 18], training the models with sufficient computational power can be done in reasonable time.

5 CONCLUSIONS AND OUTLOOK

With the rise of DL in the past decade, RNNs have become practical and powerful tools for large-scale supervised learning of sequences. This progress has become most apparent in Natural Language Processing where they have set several new benchmarks outperforming techniques that have been considered state-of-the-art for a long time. Similar to other works, the promising results from our experiments show that RNNs allow to take a novel perspective on applications such as RS that were originally designed in a time-agnostic manner. Our novel approach proposed in this paper allows generating personalized suggestions while time is considered as a first class factor, thereby significantly improving recommendation quality. The way we formulate RNNs enables us to model user behavior and to capture dependencies between consumption events more adequately than with established recommendation techniques that often fail at appropriately representing temporal dynamics in user interests.

Gated recurrent networks have set records in accuracy on many tasks in recent years. Noteworthy, these advances result from novel or extended architectures rather than from

fundamentally novel algorithms [31, 45]. This also applies to our user-based RNNs: We essentially adopted the original architecture to take the unique characteristics of the recommendation domain into account. Our specific extensions have then shown to increase statistical expressiveness: Deeply integrating a user representation leads to significant improvements when predicting user behavior such as watching movies or listening to music. Including a user-specific layer, the networks seem capable of learning concepts that are unique to a certain user. Since changing inputs are related to a user vector that remains constant over the course of a particular sequence, the networks learn to represent global user concepts explicitly, rather than implicitly as it would be the case in conventional gated units. Hence, the networks can distinguish between user preferences more accurately.

We have presented a DL-based framework that is particularly designed to generate personalized next item recommendations, thereby independent of techniques used in contemporary RS research. In future work, we plan to additionally integrate state-of-the-art recommendation techniques such as MF into the design space of RNNs, e.g. in form of contextual variables. By this means, we would be able to combine the best of both worlds while also extending the scope of applicability. Also, this might be helpful for better supporting situations where the user's consumption sequence is rather short, e.g. at cold-start. Moreover, we aim at taking varying time intervals into account. RNNs are designed for modeling sequential data with no notion of time spans between succeeding events. Especially for recommendation tasks, however, time deltas can be extremely valuable information. For instance, if two consumption events are distant from each other, the first item might not be a good predictor because user preference likely has changed over time. In this context, a comparison with statistical models particularly designed for predicting sequences would also be of interest. Finally, as interactive approaches are more and more discussed in RS research, it seems promising to examine ways of increasing control and transparency also in DL-based RS.

REFERENCES

- [1] L. Baltrunas and X. Amatriain. 2009. Towards time-dependant recommendation based on implicit feedback. In *CARS '09*.
- [2] Y. Bengio, P. Frasconi, and P. Simard. 1993. The problem of learning long-term dependencies in recurrent networks. In *ICNN '93*. IEEE, 1183–1188.
- [3] G. Bonnin and D. Jannach. 2015. Automated generation of music playlists: Survey and experiments. *ACM Comput Surv* 47, 2 (2015), 26:1–26:35.
- [4] P. G. Campos, F. Díez, and I. Cantador. 2014. Time-aware recommender systems: A comprehensive survey and analysis of existing evaluation protocols. *User Model User-Adap* 24, 1-2 (2014), 67–119.
- [5] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *SSST '14*.
- [6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Deep Learning Workshop at NIPS '14*.
- [7] Y. Ding and X. Li. 2005. Time weight collaborative filtering. In *CIKM '05*. ACM, 485–492.
- [8] S. El Hahi and Y. Bengio. 1995. Hierarchical recurrent neural networks for long-term dependencies. In *NIPS '95*. 493–499.

- [9] F. Gers. 2001. *Long short-term memory in recurrent neural networks*. Ph.D. Dissertation. University of Hannover.
- [10] I. Goodfellow, Y. Bengio, and A. Courville. 2016. *Deep learning*. MIT Press.
- [11] A. Graves. 2013. Generating sequences with recurrent neural networks. (2013).
- [12] A. Graves, A. Mohamed, and G. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *ICASSP '13*. IEEE, 6645–6649.
- [13] A. Greenstein-Messica, L. Rokach, and M. Friedman. 2017. Session-based recommendations using item embedding. In *IUI '17*. ACM, 629–633.
- [14] A. Gunawardana and G. Shani. 2009. A survey of accuracy evaluation metrics of recommendation tasks. *J Mach Learn Res* 10 (2009), 2935–2962.
- [15] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV '15*. 1026–1034.
- [16] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *CVPR '16*. IEEE, 770–778.
- [17] C. Hermann. 2010. Time-based recommendations for lecture materials. In *EdMedia '10*. 1028–1033.
- [18] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk. 2015. Session-based recommendations with recurrent neural networks. In *ICLR '16*.
- [19] B. Hidasi, M. Quadrana, A. Karatzoglou, and D. Tikk. 2016. Parallel recurrent neural network architectures for feature-rich session-based recommendations. In *RecSys '16*. ACM, 241–248.
- [20] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Comput* 9, 8 (1997), 1735–1780.
- [21] T. Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Mach Learn* 42, 1-2 (2001), 177–196.
- [22] T. Hussein, T. Linder, W. Gaulke, and J. Ziegler. 2014. Hybrid: A software framework for developing context-aware hybrid recommender systems. *User Model User-Adap* 24, 1-2 (2014), 121–174.
- [23] A. Karatzoglou. 2011. Collaborative temporal order modeling. In *RecSys '11*. ACM, 313–316.
- [24] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver. 2010. Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In *RecSys '10*. ACM, 79–86.
- [25] D. Kingma and J. Ba. 2015. Adam: A method for stochastic optimization. In *ICLR '15*.
- [26] R. Klinkenberg and T. Joachims. 2000. Detecting concept drift with support vector machines. In *ICML '00*. Morgan Kaufmann, 487–494.
- [27] Y. Koren. 2010. Collaborative filtering with temporal dynamics. *Commun ACM* 53, 4 (2010), 89–97.
- [28] Y. Koren and R. Bell. 2015. *Recommender systems handbook*. Springer, Chapter Advances in collaborative filtering, 77–118.
- [29] Y. Koren, R. Bell, and C. Volinsky. 2009. Matrix factorization techniques for recommender systems. *IEEE Computer* 42, 8 (2009), 30–37.
- [30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS '12*. 1097–1105.
- [31] Z. C. Lipton, J. Berkowitz, and C. Elkan. 2015. A critical review of recurrent neural networks for sequence learning. (2015).
- [32] T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *Workshop at ICLR '13*.
- [33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS '13*. 3111–3119.
- [34] T. Mikolov and G. Zweig. 2012. Context dependent recurrent neural network language model. In *SLT '12*. 234–239.
- [35] O. Nasraoui, J. Cerwinski, C. Rojas, and F. Gonzalez. 2007. Performance of recommendation systems in dynamic streaming environments. In *SDM '07*. SIAM, 569–574.
- [36] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio. 2014. How to construct deep recurrent neural networks. In *ICLR '14*.
- [37] D. M. Powers. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Tech* 2, 1 (2011), 37–63.
- [38] S. Rallapalli, L. Qiu, Y. Zhang, and Y.-C. Chen. 2010. Exploiting temporal stability and low-rank structure for localization in mobile networks. In *MobiCom '10*. ACM, 161–172.
- [39] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *UAI '09*. AUAI Press, 452–461.
- [40] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *WWW '10*. ACM, 811–820.
- [41] F. Ricci, L. Rokach, and B. Shapira. 2015. *Recommender systems handbook* (2nd ed.). Springer.
- [42] N. Sahoo, P. V. Singh, and T. Mukhopadhyay. 2012. A hidden Markov model for collaborative filtering. *MIS Quarterly* 36, 4 (2012), 1329–1356.
- [43] R. Salakhutdinov, A. Mnih, and G. E. Hinton. 2007. Restricted Boltzmann machines for collaborative filtering. In *ICML '07*. ACM, 791–798.
- [44] J. Schmidhuber. 1992. Learning complex, extended sequences using the principle of history compression. *Neural Comput* 4, 2 (1992), 234–242.
- [45] J. Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61 (2015), 85–117.
- [46] H. Soh, S. Sanner, M. White, and G. Jamieson. 2017. Deep sequential recommendation for personalized adaptive user interfaces. In *IUI '17*. ACM, 589–593.
- [47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *CVPR '15*. IEEE, 1–9.
- [48] Y. K. Tan, X. Xu, and Y. Liu. 2016. Improved recurrent neural networks for session-based recommendations. In *DLRS '16*. ACM, 17–22.
- [49] A. van den Oord, S. Dieleman, and B. Schrauwen. 2013. Deep content-based music recommendation. In *NIPS '13*. 2643–2651.
- [50] J. Vinagre, A. M. Jorge, and J. Gama. 2015. An overview on the exploitation of time in collaborative filtering. *Data Min Knowl Disc* 5, 5 (2015), 195–215.
- [51] E. M. Voorhees. 1999. The TREC-8 question answering track report. In *TREC '99*. 77–82.
- [52] H. Wang, N. Wang, and D.-Y. Yeung. 2015. Collaborative deep learning for recommender systems. In *KDD '15*. ACM, 1235–1244.
- [53] C. Wu, J. Wang, J. Liu, and W. Liu. 2016. Recurrent neural network based recommendation for time heterogeneous feedback. *Knowl-Based Syst* 109 (2016), 90–103.
- [54] C.-Y. Wu, A. Ahmed, A. Beutel, A. J. Smola, and H. Jing. 2017. Recurrent recommender networks. In *WSDM '17*. ACM, 495–503.
- [55] L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. Carbonell. 2010. Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *SDM '10*. SIAM, 211–222.
- [56] Z. Xu, C. Chen, T. Lukasiewicz, Y. Miao, and X. Meng. 2016. Tag-aware personalized recommendation using a deep-semantic similarity model with negative sampling. In *CIKM '16*. ACM, 1921–1924.
- [57] H.-F. Yu, N. Rao, and I. S. Dhillon. 2015. High-dimensional time series prediction with missing values. (2015).
- [58] W. Zaremba, I. Sutskever, and O. Vinyals. 2014. Recurrent neural network regularization. (2014).

The following article is reused from:

Donkers, T., Loepp, B., & Ziegler, J. (2018). Explaining Recommendations by Means of User Reviews. In A. Said, & T. Komatsu (Eds.), *Joint Proceedings of the ACM IUI 2018 Workshops* (Vol. 2068). RWTH Aachen. <http://ceur-ws.org/Vol-2068/exss8.pdf>

Explaining Recommendations by Means of User Reviews

Tim Donkers

University of Duisburg-Essen
Duisburg, Germany
tim.donkers@uni-due.de

Benedikt Loepp

University of Duisburg-Essen
Duisburg, Germany
benedikt.loepp@uni-due.de

Jürgen Ziegler

University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

The field of recommender systems has seen substantial progress in recent years in terms of algorithmic sophistication and quality of recommendations as measured by standard accuracy metrics. Yet, the systems mainly act as black boxes for the user and are limited in their capability to explain why certain items are recommended. This is particularly true when using abstract models which do not easily lend themselves for providing explanations. In many cases, however, recommendation methods are employed in scenarios where users not only rate items, but also provide feedback in the form of tags or written product reviews. Such user-generated content can serve as a useful source for deriving explanatory information that may increase the user's understanding of the underlying criteria and mechanisms that led to the results. In this paper, we describe a set of developments we undertook to couple such textual content with common recommender techniques. These developments have moved from integrating tags into collaborative filtering to employing topics and sentiments expressed in reviews to increase transparency and to give users more control over the recommendation process. Furthermore, we describe our current research goals and a first concept concerning extraction of more complex argumentative explanations from textual reviews and presenting them to users.

ACM Classification Keywords

H.3.0. Information Storage and Retrieval: General

Author Keywords

Recommender Systems; Deep Learning; Explanations

INTRODUCTION

Today's *Recommender Systems* (RS) have been shown to generate recommendations for items that match the user's interest profile quite accurately. The underlying methods are usually based either on purchase data or ratings of other users (collaborative filtering), or on structured data explicitly describing the items (content-based filtering). Yet, algorithmic maturity does not necessarily lead to a commensurate level of user satisfaction [11]. Aspects related to user experience such as the amount of control users have over the recommendation process or the transparency of the systems may also contribute substantially to the user's acceptance and appraisal of the recommendations [11]. Still, many state-of-the-art methods appear to the user as black boxes and do not provide a rationale for the recommendations, which may negatively influence intelligibility, and thus user's comprehension and trust [18].

When recommendation methods are applied in the real world, users can often provide textual feedback on items in the form of short tags or written reviews. Textual feedback from other customers is known to strongly influence the current user's decision-making [2]. However, perusing all reviews associated with the items in a recommendation set is time-consuming and mostly infeasible. The information available in review data is currently also hardly exploited for making the otherwise opaque recommendation process more transparent. While research has more recently begun to investigate the role of, for instance, product features mentioned, topics addressed, or general sentiments expressed, for improving algorithmic precision [25, 5, 1], their potential for increasing intelligibility of recommendations has not been fully exploited yet. The same applies for aspects such as presence, polarity and quality of arguments found in the reviews for or against a product. Thus, extracting and summarizing relevant arguments and presenting them as textual explanations seems to offer a promising avenue to supporting users better in their decision process.

RELATED AND PRIOR WORK

One popular way of increasing transparency of RS is to use textual explanations [20]. When sufficient content information is available, item attributes may be aligned with user preferences to explain a recommendation [22]. Such data can also be used together with context information to point out arguments for recommended items, and may simultaneously serve as a means to critique recommendations [13]. For item-based collaborative filtering, the static variant used e.g. by Amazon ("Customers who bought this item also bought...") is quite popular. For model-based methods, it is still difficult to improve transparency through explanations. The approach proposed in [25] actually exploits review data, but is limited to identifying sentiments on a phrase-level to highlight product features the user is particularly interested in. In other cases where reviews have been analyzed semantically, this has served primarily to improve model quality, e.g. by inferring hidden topics, extracting content aspects, or mining user opinions [5, 1]. More advanced approaches that, for instance, extract argumentative explanations from review texts to support the user's decision-making do not exist. One exception where argumentation techniques have been integrated into RS is presented in [4]. Yet, this approach depends on the availability of explicit item features and, since implemented in defeasible logic, requires defining postulates manually.

In prior work of ours [6, 7], we proposed a model-based recommendation method that exploits textual data. *TagMF* enhances a standard *matrix factorization* [12] algorithm with tags users

provided for the items. By learning an integrated model of ratings and tags, the meaning of the learned latent factors becomes more transparent [7] and the user may interactively change their effect on the generated recommendations [6]. In contrast to other attempts that integrate additional data, this improves comprehensibility of recommendations as well as user control over the system which is typically limited to (re-)rating single items. Moreover, our user study showed that not only objective accuracy as measured in offline experiments, but also perceived recommendation quality benefits from complementing ratings with additional tag data. Apparently, tags can introduce semantics into the underlying abstract model that are natural to understand so that users notice some kind of inner consistency in the recommendation set. Besides, the relations between users and tags introduced by our method allow to explicitly describe user preferences in textual form: We can automatically derive which tags are considered important to an individual user, even in cases when the user has never tagged items him- or herself. These tags can then be presented as an explanation of his or her formerly latent preference profile [7]. Despite its advantages, our method requires the relationship between items and additional data to be quantified in numerical form. If this requirement is not met, it seems a natural extension to exploit other, more general forms of user-generated content such as product reviews.

In [9], we presented an interactive recommending approach relying on review data. We extended the concept of *blended recommending* [15] by automatically extracting keywords from product reviews and identifying their sentiment. Then, we used the extracted (positive or negative) item descriptions to offer filtering options usually not available in contemporary RS. In our system, we present them as facet values that can be selected and weighted by the user to influence the recommendations. We conducted a user study that provided evidence that users were able to find items matching interests that are difficult to take into account when only structured content data is exploited. Without requiring users to actually read the reviews, the method seems promising for improving the recommendation process in terms of user experience, especially when users have to choose from sets of “experience products”. Although in principle any algorithm can be integrated in the system’s hybrid configuration, we have not yet utilized the advantages of model-based recommender methods in combination with the ones of exploiting user-generated reviews.

A CONCEPT FOR EXPLAINING RECOMMENDATIONS

Building on our prior work, we in the following present a novel concept that relies on extracting and summarizing arguments about products from textual reviews in order to provide users with adapted model-based recommendations, and in particular, explanations that are personalized according to the current user’s preferences and styles of decision-making.

While tagging helps to classify items by attributing specific properties, the descriptive nature of tags limits their applicability as an explanatory element. User reviews, on the other hand, constitute semantically rich information sources that incorporate sentiment and often an intended sequence of arguments, i.e. an argumentation flow. Reviews go beyond mere objec-

tive descriptions by (sometimes implicitly) invoking a stance towards a target. Users reading a review can get an idea of motives and reasoning behind its author’s words. This verbal provision of subjective information constitutes a comprehensible context on which arguments and their interrelations can be grounded. However, automatically identifying presence, polarity and quality of argumentation structures has not yet been considered in RS research, although it is well known that textual feedback of other users may strongly influence decision-making [2]. In particular, the potential of arguments has not been exploited for explaining recommendations. Due to their persuasive nature, arguments can be thought of providing intelligible reasons that support recommendations, and may thus increase system transparency and trust in the results.

Since users differ with respect to dispositions and preferences, they may attach different levels of importance to arguments. For example, one user may insist on closeness to the beach when booking a hotel, while another user lays more focus on cleanliness or friendliness of the staff. Although reading the same review, these two users would, most likely, attend to completely different aspects. A sophisticated argument extractor should therefore mimic human scanning behavior by adaptively assigning individual attention weights to arguments.

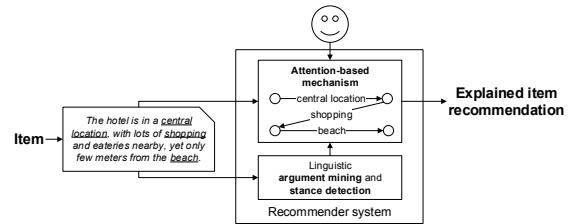


Figure 1. In our framework, a review is analyzed linguistically and via an attention-based mechanism. This allows to implement an argumentation flow based on information provided in the review while deeply integrating the user. Eventually, a personalized recommendation is presented together with individual arguments for or against this product.

We propose a conceptual framework that automates the process of extracting arguments from reviews (see Figure 1) to come up with personalized argumentative item-level explanations for recommendations: We suggest to apply feature-based methods from *computational linguistics* in order to derive argumentative structures, including argument polarity in the form of stances. Beyond that, detecting personally important arguments requires a deep integration into the process of calculating recommendations. We aim at utilizing *deep learning* methods that—matching the attention analogy—enable the system to focus on important concepts subject to a user variable. Put together, this leads to the following challenges:

- Linguistically analyzing review texts via argument mining and stance detection.
- Identifying important concepts for a target user via an attention-based mechanism.
- Deriving an argumentation flow via multiple applications of the attention-based mechanism.
- Unifying the linguistic analyses and the attention-based mechanism.

Linguistic Analyses: Computational linguistic approaches can be distinguished based on granularity of the analysis.

Document-level analyses, for instance, aim at determining a sentiment or stance towards the subject of decision, e.g. a hotel. However, for our purpose such an approach is too shallow as a review generally not only consists of utterances regarding the target item as a whole, but usually also includes remarks on sub-aspects. Sentiments or stances towards these sub-aspects may deviate from the review’s overall polarity. For example, a guest may generally like a hotel albeit he or she found the bed uncomfortable. In addition, a review might address sub-aspects not important to the current user. Therefore, we propose a more fine-grained approach relying on aspect-level argument mining that is capable of assigning polarity to a variety of mentioned entities.

Automated argument mining refers to identifying linguistic structures consisting of at least one explicit claim and optional supporting structures such as premises [17]. Depending on which theory of argumentation one follows [21, 10], these structures are more or less specialized. However, analyzing user-generated reviews depicts a difficult task for an argument mining tool as they deviate considerably from texts on which linguistic argument mining is typically performed, e.g. legal text or scientific writing: Reviews are usually shorter, noisier, less densely packed with arguments, and contain arguments in a way that is often not as sophisticated.

Another difficulty arises when one wants to decide whether an extracted argument is in favor or against a particular product. However, a notion of polarity is necessary to be able to select adequate arguments supporting or opposing a recommendation. Therefore, we propose to relate each argument to a stance expressed by the review author. In this regard, a stance target is not limited to the subject of decision itself but may essentially address anything towards which one can have a stance. It follows that the linguistic analyses need to identify the target and establish a relation to the subject of decision. Please note that stance detection involves identifying a subjective disposition that might often be implicit [14]. Moreover, stances, as opposed to e.g. sentiments, do not necessarily integrate a polarity aspect (e.g. “the hotel is modern”, “the food is local”).

Attention-based Mechanism: Since we are interested in providing users with personalized explanations, solely performing linguistic analyses is not sufficient: Users differ not only in their interests, but have unique styles of decision-making. As a consequence, assessment of the importance of an argument found in a review as well as its accordance with the stances expressed by the review author have to be adapted towards the current user. In order to achieve this, we propose an attention-based approach that considers a vectorial user representation to identify personally important concepts.

Attention-based deep learning approaches [3] have proved to be very successful at identifying significant local features in tasks from several domains such as image processing [24] or machine translation [16]. An attention function can be described as a (soft-)search for a set of positions in a linguistic source where the most relevant information is concentrated. In the context of review-based RS, we propose to use attention as a means to identify important and distinguishing concepts for the subject of decision, i.e. a target item. Technically speaking,

we suggest to compute a weighted sum over the sequence of vectors derived from the words contained in a review. The assigned weights would indicate the relative importance of each vector and thus resemble the amount of attention a particular word is receiving.

Personalizing explanations requires attention to be distributed with respect to user preferences, i.e. users act as the context that moderates the (soft-)attention’s output. Thus, it is necessary to calculate attention weights subject to a vectorial user representation. We propose to model users analogous to our previous work [8] by embedding one-hot user vectors, along with word vectors, into a densified joint information space. This would allow to numerically estimate the degree to which the current user’s preferences are in line with concepts expressed in a review. Assume, for example, the following portion of a review: “The food is good, but beds are uncomfortable.” If the user has shown interest in good food in the past, the attention mechanism should assign a large weight to *food*. Although the lack of *bed* comfort might be relevant to others, this argument should be neglected and receive a weight close to zero if the system has not detected a relationship between the current user and this particular aspect before.

Argumentative Flow: Argumentation can be considered more convincing if it consists not only of an aggregation of potentially important words or phrases. We assume arguments to become more understandable and effective in case they follow a coherent argumentative flow, i.e. dependencies between successively extracted arguments. The analogy in our proposed concept is the repeated application of the (soft-)attention mechanism while considering previous output as context. Informally, this mimics a conversational exchange between user and attention mechanism where the latter continuously tries to identify concepts more and more tailored specifically towards the user’s preferences. The result would be a succession of relevant word sets that exhibit sequential properties and are thus tied to each other. As a consequence, the output sequence would reflect the whole path of how the system came up with a particular recommendation and thus explain which concepts played an important role during the process (see Figure 1). It must be noted that such a procedure would be closely related to so-called *memory networks* [23, 19], which also work with multiple hop operations on an attention-based memory. Since our concept imposes an artificial argumentative structure on the raw review text, it will be interesting to investigate the argumentative flow intended by the review’s author compared to the one automatically derived by the system.

Unification of Linguistic Analyses and Attention: Up to this point, we have covered two independent approaches to process review texts: (1) linguistic analyses aiming at mining arguments and detecting stances, and (2) attention-based extraction of words relevant to the current user. However, both on their own are limited in expressiveness: the linguistic analyses lack the personalization component while the attention mechanism operates on word-level, thus incapable of extracting complete arguments. Consequently, following our superordinate research goal of presenting users more informative, personalized explanations, we plan to exploit the benefits

of both approaches by coupling them closely together. A first attempt, for instance, would be to check whether the surrounding context of a word that received a large amount of attention was identified as an argument by the linguistic processor. If this condition is met, the system can interpret the argument as a possible candidate for a personalized explanation. Assume the system detects e.g. the arguments “the food is good” and “beds are uncomfortable”, then identifying *food* as an important individual concept should lead to the first argument being chosen for explaining the recommendation.

CONCLUSIONS

In this paper, we have discussed the state of research regarding usage of product reviews in RS—with a focus on explanations. We set our prior work into context, where we already used tags to increase a recommender’s transparency and analyzed product reviews in order to provide extended interaction possibilities. Based on this, we pointed out a possible way to exploit this rich information source for presenting users with intelligible, personalized explanations by extracting more complex arguments. We outlined several challenges and proposed a concept to address them. For future work, our research goals are to implement this novel concept, use it to additionally personalize recommendations, and to evaluate it in several domains with a focus on the influence on user experience in comparison to other explanation methods.

REFERENCES

1. A. Almahairi, K. Kastner, K. Cho, and A. Courville. 2015. Learning Distributed Representations from Reviews for Collaborative Filtering. In *RecSys '15*. ACM, 147–154.
2. G. Askalidis and E. C. Malthouse. 2016. The Value of Online Customer Reviews. In *RecSys '16*. ACM, 155–158.
3. D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR '15*.
4. C. E. Briguez, M. C. D. Budán, C. A. D. Deagustini, A. G. Maguitman, M. Capobianco, and G. R. Simari. 2014. Argument-based Mixed Recommenders and Their Application to Movie Suggestion. *Expert Syst. Appl.* 41, 14 (2014), 6467–6482.
5. Q. Diao, M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, and C. Wang. 2014. Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS). In *KDD '14*. ACM, 193–202.
6. T. Donkers, B. Loepp, and J. Ziegler. 2016a. Tag-Enhanced Collaborative Filtering for Increasing Transparency and Interactive Control. In *UMAP '16*. ACM, 169–173.
7. T. Donkers, B. Loepp, and J. Ziegler. 2016b. Towards Understanding Latent Factors and User Profiles by Enhancing Matrix Factorization with Tags. In *RecSys '16*.
8. T. Donkers, B. Loepp, and J. Ziegler. 2017. Sequential User-Based Recurrent Neural Network Recommendations. In *RecSys '17*. ACM, 152–160.
9. J. Feuerbach, B. Loepp, C.-M. Barbu, and J. Ziegler. 2017. Enhancing an Interactive Recommendation System with Review-based Information Filtering. In *IntRS '17*. 2–9.
10. J. B. Freeman. 1991. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. De Gruyter.
11. J. A. Konstan and J. Riedl. 2012. Recommender Systems: From Algorithms to User Experience. *User Mod. User-Adap.* 22, 1-2 (2012), 101–123.
12. Y. Koren, R. M. Bell, and C. Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer* 42, 8 (2009), 30–37.
13. B. Lamche, U. Adıgüzel, and W. Wörndl. 2014. Interactive Explanations in Mobile Shopping Recommender Systems. In *IntRS '14*. 14–21.
14. W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. 2006. Which Side are you on?: Identifying Perspectives at the Document and Sentence Levels. In *CoNLL-X '06*. Association for Computational Linguistics, 109–116.
15. B. Loepp, K. Herrmann, and J. Ziegler. 2015. Blended Recommending: Integrating Interactive Information Filtering and Algorithmic Recommender Techniques. In *CHI '15*. ACM, 975–984.
16. T. Luong, H. Pham, and C. D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *EMNLP '15*. 1412–1421.
17. R. M. Palau and M.-F. Moens. 2009. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *ICAIL '09*. ACM, 98–107.
18. P. Pu, L. Chen, and R. Hu. 2012. Evaluating Recommender Systems from the User’s Perspective: Survey of the State of the Art. *User Mod. User-Adap.* 22, 4-5 (2012), 317–355.
19. S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. 2015. End-To-End Memory Networks. In *NIPS '15*. 2440–2448.
20. N. Tintarev and J. Masthoff. 2015. *Recommender Systems Handbook*. Springer US, Chapter Explaining Recommendations: Design and Evaluation, 353–382.
21. S. E. Toulmin. 2003. *The Uses of Argument*. Cambridge University Press.
22. J. Vig, S. Sen, and J. Riedl. 2009. Tagsplanations: Explaining Recommendations Using Tags. In *IUI '09*. ACM, 47–56.
23. J. Weston, S. Chopra, and A. Bordes. 2014. Memory Networks. (2014).
24. K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML '15*. 2048–2057.
25. Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma. 2014. Explicit Factor Models for Explainable Recommendation Based on Phrase-Level Sentiment Analysis. In *SIGIR '14*. ACM, 83–92.

The following article is reused from:

Loepp, B., Donkers, T., Kleemann, T., & Ziegler, J. (2019). Interactive recommending with Tag-Enhanced Matrix Factorization (TagMF). *International Journal of Human-Computer Studies*, 121, 21–41. <https://doi.org/10.1016/j.ijhcs.2018.05.002>

Interactive Recommending with Tag-Enhanced Matrix Factorization (*TagMF*)[☆]

Benedikt Loepp*, Tim Donkers, Timm Kleemann, Jürgen Ziegler

University of Duisburg-Essen, Duisburg, Germany

Abstract

We introduce *TagMF*, a model-based Collaborative Filtering method that aims at increasing transparency and offering richer interaction possibilities in current Recommender Systems. Model-based Collaborative Filtering is currently the most popular method that predominantly uses Matrix Factorization: This technique achieves high accuracy in recommending interesting items to individual users by learning latent factors from implicit feedback or ratings the community of users provided for the items. However, the model learned and the resulting recommendations can neither be explained, nor can users be enabled to influence the recommendation process except by rating (more) items. In *TagMF*, we enhance a latent factor model with additional content information, specifically tags users provided for the items. The main contributions of our method are to use this integrated model to elucidate the hidden semantics of the latent factors and to let users interactively control recommendations by changing the influence of the factors through easily comprehensible tags: Users can express their interests, interactively manipulate results, and critique recommended items—at cold-start when no historical data is yet available for a new user, as well as in case a long-term profile representing the current user’s preferences already exists.

To validate our method, we performed offline experiments and conducted two empirical user studies where we compared a recommender that employs *TagMF* against two established baselines, standard Matrix Factorization based on ratings, and a purely tag-based interactive approach. This user-centric evaluation confirmed that enhancing a model-based method with additional information positively affects perceived recommendation quality. Moreover, recommendations were considered more transparent and users were more satisfied with their final choice. Overall, learning an integrated model and implementing the interactive features that become possible as an extension to contemporary systems with *TagMF* appears beneficial for the subjective assessment of several system aspects, the level of control users are able to exert over the recommendation process, as well as user experience in general.

Keywords: Recommender Systems, Collaborative Filtering, Interactive Recommending, Matrix Factorization, Tags, Empirical Studies, Human Factors, User Experience, User Interfaces, User Profiles

1. Introduction

Recommender Systems (RS) based on *Collaborative Filtering* (CF) have been shown to be effective means for leveraging the “wisdom of the crowd” to identify items that are potentially of interest to a user. They support users in finding items that match their personal preferences from very large sets of items, such as, for instance, consumer goods, documents, or movies [1, 2]. From an information provider’s perspective, a major advantage of CF recommenders lies in the fact that only feedback the community of users provided for the items—explicitly expressed via ratings or implicitly acquired through user actions—is required as input data [3]. Considerable advances have been made in recent years with respect to the objective performance of CF systems as measured by common accuracy

metrics in retrospective offline experiments [4]. However, it has been observed that high offline recommendation accuracy (i.e. accurately predicting which items should be recommended to a user) does not necessarily lead to a commensurate level of user satisfaction [5, 6, 7]. Since CF algorithms are considered already quite mature, the small incremental improvements that still seem possible with respect to algorithmic precision are thus not likely to be particularly beneficial for users. Consequently, other evaluation metrics have been discussed to assess the quality of recommendation sets, for example, diversity, novelty, and serendipity [8, 9]. Beyond that, one important aspect that may contribute to actual user satisfaction is the degree of control users have over the systems [6, 10]. Yet, from a user’s perspective, the ways to influence the generation of recommendations in today’s automated RS such as the ones used by Amazon [11] or Netflix [12] are mostly very limited. Usually, the only means to actively influence the results is to provide explicit feedback about the items’ relevance, i.e. rating or re-rating single items. Among others, this raises the risk of users being stuck in a “filter bubble” [13] as the recommendations are increas-

[☆] © 2019. This manuscript version is made available under the CC-BY-NC-ND 4.0 license. DOI: 10.1016/j.ijhcs.2018.05.002.

*Corresponding author

Email addresses: benedikt.loeppl@uni-due.de (Benedikt Loepp), tim.donkers@uni-due.de (Tim Donkers), timm.kleemann@uni-due.de (Timm Kleemann), juergen.ziegler@uni-due.de (Jürgen Ziegler)

ingly constrained to items similar to those the current user has rated positively in the past. This well-known effect makes it difficult to become aware of hidden alternatives, to explore new and diverse areas of potential interest, and to adapt the results towards situational needs and goals [13, 14]. A further problem can be seen in the general lack of transparency of contemporary CF recommenders [5, 7]. The methods prevalently used infer abstract models from the original input data, making it difficult for users to understand the profile that represents their preferences, and consequently why certain items are recommended. This, in turn, may reduce user trust in the system as well as acceptance of suggested items [5, 15]. Hence, adding more interactivity to the system and letting users influence the recommendation process as well as making it more comprehensible is increasingly considered an important goal in RS research [5, 6, 7, 16, 10].

Only more recently, such aspects related to user experience of RS have begun to attract more attention [6, 17]. In this line of research, interactive recommenders have been proposed that use, for instance, metadata such as user-generated tags to calculate recommendations and to offer users additional interaction mechanisms [18]. Using tags has the advantage of relying on concepts that are meaningful to users without requiring explicit item descriptions. Consequently, eliciting preferences via tags has been shown to bear the potential for improving user control and comprehension [19]. However, tag-based RS [e.g. 19, 20, 18] have typically been developed independently of established CF methods. For this reason, such systems cannot benefit from existing long-term preference profiles based on implicit feedback or rating data. The same applies for most of the recommending approaches that aim at increasing interactivity in general [e.g. 21, 22, 23]. In particular, they usually do not exploit model-based CF techniques such as the widely used *Matrix Factorization* (MF) [24, 25], which is known for efficiency and has been shown to achieve high offline accuracy. On the other hand, models as derived by MF have only rarely been applied for purposes other than improving recommendation effectiveness or algorithmic performance. What is lacking, therefore, are methods that combine the accuracy-related benefits of model-based CF with the easy-to-understand semantics of user-generated tags.

In this paper, we introduce *Tag-Enhanced Matrix Factorization* (*TagMF*), a novel recommendation method that enhances a MF model with tags that users provided for the items, and propose several possible applications of *TagMF* for realizing interactive recommenders by extending conventional model-based CF systems. MF models represent both users and items in a joint latent factor space [24]. Since latent factors are usually learned by statistically analyzing historical implicit feedback or rating data, the semantics of these factors are hidden, yet are generally considered to relate to real-world concepts [24, 26]. For instance, the factors may describe more or less obvious characteristics such as the “amount of action” a user

appreciates or the “degree of black humor” in a movie. Once the factors have been learned, latent factor models allow to accurately predict ratings for items the user has not yet seen, or to establish a ranking among them. Several approaches already exist that employ additional information such as context data, predefined or user-generated content-related metadata, or topics and opinions inferred from user reviews [e.g. 27, 28, 29, 30, 31, 32, 33, 34]. The respective methods have been applied with the goal of further improving model quality, and as a consequence offline recommendation accuracy (i.e. how well the predictions match implicit feedback or ratings provided in the past), not for exposing the additional information at the user interface. Accordingly, there are currently also no user studies that show the benefits of enhancing a model-based CF recommender with additional content information. With *TagMF*, we contribute to the state of research by answering the following research questions:

RQ1: How can additional information be used in model-based CF systems for ...

- a) eliciting preferences in cold-start situations without requiring the user to rate items?
- b) manipulating recommendations resulting from an existing user profile?
- c) critiquing a recommended item while considering the user’s long-term interests?
- d) explaining an existing preference profile?

RQ2: How does additional information affect subjective system aspects such as perceived recommendation quality and user experience when compared to ...

- a) an automated recommender based on ratings?
- b) an interactive recommender based on tags?

While we use user-generated tags as additional content information in this paper, our algorithmic method can in principle be applied to any other type of descriptive item information. We aim at showing that enhancing MF is not only beneficial in terms of model quality, but also with respect to user experience. By employing *TagMF*, users can interactively express their preferences and control the recommendation process in a model-based CF recommender via tags. While ratings stored in an existing user profile or provided during interaction are still taken into account, users can by this means indirectly determine their preferences in the space spanned by the latent factors and interactively adapt the set of recommendations without being required to (re-)rate items. This is possible both in cold-start situations, i.e. for new users entering the system who do not yet have an existing long-term profile, as well as when a profile based on past user feedback is persistently available in the system but the user’s needs deviate from long-term interests. Availability of the current user’s rating data is not mandatory. Instead, our method requires as input only a conventional dataset of implicit feedback or ratings (of other users) as well as item-related tag relevance

information. From this point of departure, the method allows to derive user-related tag relevance information as well as tag-factor relations. Thus, users themselves do not need to have tagged items before, i.e. we do not require to know a priori how relevant tags are for the current user as we infer this information. Moreover, integrating the easy-to-understand semantics of tags in this novel way allows us to open up the “black box” latent factor models usually constitute for the user. With *TagMF*, we are able to establish a general understanding of the factor space, and to show how users and items are positioned inside it. As a consequence, users can be presented with explicit tag-based explanations of their profile representing preferences they have expressed indirectly with respect to the nontransparent factor space.

To evaluate our method, we first conducted extensive offline experiments comprising an analysis of objective performance and a qualitative inspection of a resulting factor model. Then, in order to validate the application possibilities of *TagMF* and to examine user experience, we implemented a web-based prototype movie RS that uses our method for generating recommendations and for providing users with additional tag-based interaction possibilities. In two quantitative user studies, we compared this interactive system both with a conventional automated recommender that uses MF and with a purely tag-based interactive approach. To the best of our knowledge, our evaluation hence forms the first and most extensive empirical examination of the effects considering additional information has on CF recommenders to date. Among several promising findings, the results indicate that learning an integrated model increases perceived recommendation quality, which previously has only been observed in offline experiments [e.g. 27, 29, 30, 35, 32, 34]. To further analyze aspects related to user experience, we used *Structural Equation Modeling* (SEM) [36]. SEM, a multivariate analysis technique which is still rarely applied in RS research [17, 16], allowed us to investigate the influence applying our method has on the measurement of such aspects and the relationships between them. The analysis yields interesting insights, among others, that users perceive recommendations to be more transparent, and are as a consequence more satisfied with the item finally chosen, when they can additionally interact via tags. In general, the results emphasize the value of considering latent knowledge and (user-generated) content information at the same time—both for improving recommendations and extending interactive control in contemporary RS.

In the following, we first discuss relevant related work. Next, we describe the methodology behind *TagMF* in detail, and elaborate on its application possibilities that allow to implement interactive RS. Afterwards, we present our evaluation, including offline experiments and user studies. Finally, we conclude the paper by discussing the results and providing an outlook on future work.

2. Related Work

Successful examples of commercial recommenders are the systems used by Amazon [11] or Netflix [12], which aim at presenting recommendations that fit well the user’s general preferences while reducing interaction effort and cognitive load. However, users might feel too much dominated by the systems, unable to flexibly specify current interests or to obtain, for instance, more diverse and novel recommendations. This is particularly true because users are mostly very limited in their ways to interact with such automated RS or have no control over the recommendation process at all, although this might considerably increase user satisfaction [5, 6, 7, 10]. In contemporary CF recommenders, the only way for users to actively affect the results is usually by providing explicit feedback in form of ratings for single items. While this represents a possibility to exert at least some influence, it does not eliminate the “filter bubble” effect [13] since the user’s existing long-term profile is only further refined despite the fact that the search goal may vary depending on the current situation. Moreover, considerable effort is required on part of the user before he or she can obtain adequate suggestions [37, 38]—especially in cold-start situations, i.e. when no historical data is yet available for a new user entering the system or when a user does not want an existing profile to be applied. Apart from that, notably in real-world systems, results are often adapted based on implicit feedback, for example, when users click on interesting items to see more details [3, 2, 10]. This way, user interaction behavior can be modeled more accurately compared to ratings [39, 3], but the process tends to become less transparent and it gets even harder for users to adapt the recommendations towards their situational needs.

2.1. Providing Control and Improving Transparency

In light of these drawbacks, interactive approaches that focus on increasing the level of user control over the recommendation process and improving its transparency have received more and more attention in recent years [40, 41, 10]. For instance, in critique-based RS, users can manipulate the results by critiquing a suggested item with respect to product properties they wish to value higher or lower [21]. In contrast to such attempts, *MovieTuner* [18] does not require previously modeled metadata. Instead, it solely relies on user-generated tags allowing to ask for a movie similar to the currently recommended one—but e.g. less violent and more funny. For tailoring the critiquing process towards the current user, past critiquing sessions can be taken into account [42]. However, long-term profiles as they are customary in CF systems are usually neither considered for adapting the process itself nor do they eventually affect the recommendations.

TasteWeights [22], *SetFusion* [23], *MyMovieMixer* [43] and *uRank* [44] allow to control a RS in a more advanced manner: Users can interactively vary the influence of different social datasources [22], of various algorithms [23], of

certain product facets [43], or of extracted keywords [44] in order to better reflect their current interests. Moreover, these approaches aim at improving system transparency through visualizations that support users in understanding why the items were recommended. Related examples which make even more extensive use of visualization techniques comprise, among others, *MoodPlay* [45] and *Conference Navigator* [46]. A comprehensive overview including attempts to visualize item space and user profiles can be found in [40, 10].

While the most popular type of recommender algorithms is CF [2], many of the attempts proposed to increase interactivity and transparency, including the ones mentioned above, are developed independently of CF: They typically rely on their own concepts to recommend items instead of building on the benefits of established model-based CF techniques that are known for high precision and efficiency [25]. Consequently, even when available, past browsing behavior or previously given ratings cannot be taken into account. Against this background, to our knowledge, no attempts have been made to extend a model-based CF recommender into a fully interactive, user-controlled system.

2.2. Extending Matrix Factorization

Despite the success of MF techniques that learn latent factor models, RS research has been trying to further increase recommendation quality in terms of objective performance metrics [4, 16]. One promising attempt is to complement existing ratings with further data. This additional information may be rather generic, such as implicit user feedback or temporal relations of ratings [24, 25], but often, more specific datasources are taken into account: In [28, 35], predefined content-related metadata about movie genres or recipe ingredients are exploited. Other approaches rely on contextual information, for example, user age or current season [e.g. 27]. Several authors semantically analyze user-written product reviews to first infer hidden topics or opinions about the items, which are subsequently integrated with latent factor models to improve their quality [e.g. 29, 31, 33, 34]. However, only few approaches take immediate advantage of user-generated information such as tags. In these approaches, the underlying models are enhanced with, for instance, specific keywords regarding a movie’s mood and plot [30] or generic social tags [32], but are focused on improving offline accuracy rather than user control and system transparency. Accordingly, they have not been evaluated in user studies, leaving the influence on user experience open for investigation. Besides, there exist indeed approaches that rely exclusively on tags for the purpose of generating recommendations, e.g. using graph-based methods [20] or by directly modeling user preferences based on item-tag signals [19, 47]. Yet, they are limited since they cannot benefit from the algorithmic maturity of model-based CF techniques, and thus the availability of existing long-term

preference profiles based on implicit or explicit user feedback data. Moreover, apart from e.g. *MovieTuner* [18] or *uRank* [44], these tag-based RS are again not particularly designed for giving users more interactive control.

The range of techniques for considering additional information in CF recommenders is very broad as well. For standard MF [24], which is closely related to *Singular Value Decomposition* (SVD) [48] and thus often referred to as “SVD-like MF”, not needing imputation and preventing overfitting by means of regularization [49, 25], a straight-forward way is to add further constraints to the minimization function that is used to learn the parameters when training the latent factor model. This increases precision [24, 32], but after having been learned, the latent factors exhibit no interpretable association with the additional information: The information is calculated into the factor values in a way that the relationship between provided data and latent factors, and consequently items, cannot be made accessible for users anymore. The same applies to approaches that use additional regularization terms [e.g. 29, 30]. In contrast, in [28, 35, 33], the information is explicitly used to establish a content-related association with the factors: By proposing a regression-constrained formulation, factors are considered as functions of content attributes. Further techniques for enriching model-based CF are, among others, extended probabilistic MF [31], deep learning [34, 50], factorization machines [51], or the generalized variant of MF, tensor factorization [27]. All of these attempts have been shown to significantly increase the accuracy objectively measurable in offline experiments. However, to our knowledge, there are currently no empirical user studies available that examine the effects of integrating CF, and in particular latent factor models, with additional information in terms of subjective aspects such as perceived recommendation quality and variety, or user experience in general.

2.3. Exploiting Latent Factor Models

Overall, the usage of latent factor models has rarely been exploited for purposes other than improving effectiveness or performance of RS. Nonetheless, while cold-start situations in CF have mostly been addressed algorithmically [e.g. 52, 53, 54, 55], some exceptions rely on the factor space to interactively elicit initial user preferences, for instance, in a choice-based manner [56, 57]. Likewise, latent factors may contribute to diversify a recommender’s output [e.g. 58]. Moreover, notably without the need for explicit content information, they can provide a basis to visualize the item space, e.g. in form of a map [59]. Recently, this metaphor has been extended to a 3D landscape, where the additional dimension represents the current user’s interests and allows to interactively express preferences, both with and without an existing profile [60].

In cases where MF has been actually enhanced with additional information as described in Section 2.2, this has primarily served to improve accuracy, not for exposing the additional content information at the user interface. One

of the few exceptions is [61], where user and item characteristics are explained by visualizing the importance of tags according to their correlation with the factors. In [62], a first step towards automatically explaining latent factors in textual form has been taken by associating them with topics inferred from unstructured data. Nevertheless, factors as derived by MF can still be considered overall hard to explain due to their statistical nature, and it seems particularly difficult from a system perspective to relate them to an intelligible meaning [24]. Besides, it can be seen as a more fundamental problem of model-based CF that users typically lack deeper understanding of the underlying mechanisms [5, 7, 15]. For these reasons, latent factor models have yet only rarely been suggested as a means to improve interactive control and transparency in RS.

2.4. Evaluating Recommender Systems

While aspects related to user experience are increasingly considered important for RS research [6, 7, 17], still only few evaluations go beyond measuring performance in retrospective offline experiments [16]. Especially recommenders enhanced with additional information such as tags have not yet been extensively analyzed in empirical user studies. In order to evaluate the user’s perception of system and recommendations, the framework proposed in [17] constitutes an important means to explain, among others, how subjective system aspects (e.g. perceived recommendation quality) mediate the impact of objective system aspects (e.g. differences in algorithms) on user experience. Advanced multivariate analysis techniques such as SEM that allow to investigate the underlying relationships are however only rarely used in RS research although they have been considered particularly useful for evaluating user experience [17, 16]. Exceptions have analyzed, for instance, effects of objective system aspects on perception of results [17, 63], influence of choice-based preference elicitation compared to a conventional rating phase [57], how the number of recommended items affects choice difficulty and satisfaction [64], and how diversification based on latent factors may improve these aspects [58]. As already pointed out, it has however not yet been empirically examined how considering additional information in model-based CF actually influences user experience.

2.5. Summary

In summary, it seems promising to extend MF in a way that latent factors can be associated with concepts users understand. Consequently, users could be enabled to interactively control the recommendation process according to their situational needs—both in cold-start situations as well as with an existing preference profile—and be presented with explanations of their formerly opaque representation within the factor model. In this regard, it appears of particular interest to investigate the impact on the subjective assessment of system aspects such as recommendation quality and on user experience in general.

3. Methodology

In this section, we describe *TagMF*, a method to enhance a model-based CF recommender that relies on common user-item interaction data, i.e. implicit feedback or explicit ratings users provided for the items, with additional content information, specifically tags assigned to items by the user community. We show how to learn a model that integrates this item-related tag relevance information in order to subsequently derive corresponding user-tag relevance scores as well as tag-factor relations¹.

In CF, user-item interaction data is usually represented by means of a typically sparse user-item matrix $\mathbf{R} \in \mathbb{R}^{|U| \times |I|}$. By conventional notation, each entry of \mathbf{R} represents a rating r_{ui} given by user $u \in U$ to item $i \in I$, where U is the set of users and I is the set of items [2, 25]. Note that possible values for r_{ui} may differ depending on the application: Typically, the values are numerical ratings (e.g. 1–5), but \mathbf{R} may also contain binary implicit feedback data.

Standard SVD-like MF (see Section 2.2) reduces the dimensionality of \mathbf{R} by learning a latent factor model which then serves to generate recommendations [24, 25]. This model approximates \mathbf{R} through two low-rank matrices, $\mathbf{P} \in \mathbb{R}^{|U| \times |F|}$ and $\mathbf{Q} \in \mathbb{R}^{|I| \times |F|}$, where F is a set of latent factors². The user-factor matrix \mathbf{P} and the item-factor matrix \mathbf{Q} can be trained using optimization algorithms such as Stochastic Gradient Descent or Alternating Least Squares, which are able to efficiently handle sparse matrices [24, 25]. A user’s u (calculated) interest in a particular factor f is then numerically expressed by entry p_{uf} of \mathbf{P} while entry q_{if} of \mathbf{Q} describes the extent to which item i possesses this factor. Consequently, with users and items being mapped into the same factor space [24], the inner product of a user-factor vector p_u and an item-factor vector q_i captures the interaction between user and item, and thus allows to predict the rating \hat{r}_{ui} of user u for item i . Overall, this results in:

$$\mathbf{R} \approx \mathbf{P}\mathbf{Q}^T \quad (1)$$

As our method is in principle independent of algorithmic details, we omit elaborating on e.g. regularization terms and refer to the literature [e.g. 24, 25] for a general and more extensive introduction to MF.

3.1. Integrating Item-Related Tag Information

Since a latent factor model derived as described above cannot be directly integrated with additional information, we need to add further constraints. Initially following the approach proposed in [28] (see Section 2.2), we complement a SVD-like MF algorithm by extending \mathbf{Q} with item-related information. For this, we use tag relevance scores

¹The basic principle of our method was introduced in the poster publication of [65]. Now, we describe the method in more detail and subsequently discuss the possibilities to actually apply *TagMF* in common model-based CF systems.

²The number of factors (typically 10 to 100) has to be specified before the actual factorization.

for items relying on a set of tags T , and define $\mathbf{A} \in \mathbb{R}^{|I| \times |T|}$ as a matrix representing how strongly items relate to tags: Each entry a_{it} of \mathbf{A} describes on a continuous scale from 0 (not relevant) to 1 (very relevant) the degree to which a tag t is relevant for an item i . However, we additionally extend \mathbf{P} and define $\mathbf{A} \in \mathbb{R}^{|U| \times |T|}$ to also represent user-tag relations, i.e. tag relevance scores for users. From that, we redefine the original MF model given in (1) as follows:

$$\mathbf{R} \approx \mathbf{P}\mathbf{Q}^T = \mathbf{A}\mathbf{U}\Theta(\mathbf{A}\mathbf{I}\Theta)^T, \quad (2)$$

where $\mathbf{U}\Theta \in \mathbb{R}^{|T| \times |F|}$ associates tags with factors as seen from user side and $\mathbf{I}\Theta \in \mathbb{R}^{|T| \times |F|}$ is the equivalent for items. In fact, this represents a regression-constrained formulation of the MF problem, where each factor is a function of the content attributes.

Additional content information may only be available either for users or for items. In [28], for instance, content-related metadata explicitly defined for the items has been taken into account (see Section 2.2). Since we aim at enhancing MF with tags users provided for the items, we assume that item-related tag relevance information is known a priori, and the corresponding matrix \mathbf{A} has been determined separately with a suitable method. In principle, this relationship between items and additional information can be quantified using any type of attribute that relates to both user information space and item information space in a meaningful way. The only requirement is that a numerical representation can be derived so that the entries of \mathbf{A} hold the respective relevance scores for items on a continuous scale. Information on which specific users applied which tags is however not required a priori: In contrast to matrix \mathbf{A} , we consider the corresponding matrix for users, \mathbf{A} , to be unknown. Consequently, we treat the whole term $\mathbf{A}\mathbf{U}\Theta$ implicitly at this step by just learning the user-factor matrix \mathbf{P} as known from standard MF. With this constrained equation, we can now formulate the following minimization problem as done in [28]:

$$\min_{\mathbf{P}, \mathbf{I}\Theta} \sum_{(u,i) \in K} (r_{ui} - p_u^T \mathbf{I}\Theta^T a_i)^2 + \lambda \left(\sum_{u \in U} \|p_u\|^2 + \|\mathbf{I}\Theta\|^2 \right), \quad (3)$$

with λ controlling the extent of regularization and K being the set of all user-item tuples for which user feedback (e.g. ratings) exists. We then apply a gradient descent algorithm with learning rate μ to minimize the error:

$$\begin{aligned} p_u &\leftarrow p_u + \mu \left(\sum_{i \in K_u} (r_{ui} - p_u^T \mathbf{I}\Theta^T a_i) \mathbf{I}\Theta^T a_i - \lambda p_u \right) \\ \mathbf{I}\Theta &\leftarrow \mathbf{I}\Theta + \mu \left(\sum_{(u,i) \in K} (r_{ui} - p_u^T \mathbf{I}\Theta^T a_i) a_i p_u^T - \lambda \mathbf{I}\Theta \right) \end{aligned} \quad (4)$$

3.2. Deriving Tag-Factor Relations for Users

At this point, we have transferred the abstract factor semantics into a comprehensible information space utilizing a regression-constrained approach on the item side. Although we considered tag relevance scores to be known

only for items, we can now establish a relationship between users and tags, enabling us later to let users specify their interests via tags and to explain their profile to them.

For this purpose, we apply the learned relationship between tags and latent factors to the user side. This is possible as the way a MF model is learned (see above) ensures per definition that both users and items are mapped into a joint factor space [24]. Thus, each factor $f \in F$ reflects a certain characteristic that has the same (hidden) semantic meaning for both users and items [24, 26]. The regression coefficients hence describe tag-factor relations in general, for users as well as for items. Accordingly, the implicitly assumed $\mathbf{U}\Theta$ is equivalent to $\mathbf{I}\Theta$, such that:

$$\mathbf{U}\Theta = \mathbf{I}\Theta =: \Theta \quad (5)$$

As a consequence, \mathbf{A} is now the only unknown left. Based on our problem formulation, its row vectors a_u should hold the equivalents of the item-related tag relevance scores from \mathbf{A} with respect to users. In accordance with (2), we thus solve for \mathbf{A} :

$$\begin{aligned} \mathbf{P} &= \mathbf{A}\Theta && \Leftrightarrow \\ \mathbf{P} &= \mathbf{A}\mathbf{U}\Sigma\mathbf{V}^T && \Leftrightarrow \\ \mathbf{A} &= \mathbf{P}\mathbf{V}\Sigma^+\mathbf{U}^T && \Leftrightarrow \\ \mathbf{A} &= \mathbf{P}\Theta^+ \end{aligned} \quad (6)$$

Since Θ is generally not a square matrix, we have to calculate its pseudoinverse Θ^+ (i.e. the Moore-Penrose generalization of the inverse matrix [66, 67]) first by applying SVD [48], yielding $\mathbf{U} \in \mathbb{R}^{|T| \times |T|}$, $\Sigma \in \mathbb{R}^{|T| \times |F|}$ and $\mathbf{V} \in \mathbb{R}^{|F| \times |F|}$. Consequently, Θ^+ is defined as $\mathbf{V}\Sigma^+\mathbf{U}^T$.

The general interest of user u with respect to all tags provided by the user community is now expressed by vector a_u of \mathbf{A} , which is easy to understand and basically the calculated counterpart of the given item-tag relevance scores introduced in Section 3.1.

Finally, since $\Theta\Theta^T$ holding the general tag-factor relations in (2) is a square diagonalizable matrix, we can represent it in terms of eigenvalues and eigenvectors using eigendecomposition:

$$\begin{aligned} \mathbf{R} &\approx \mathbf{A}\Theta\Theta^T\mathbf{A}^T \\ &\approx \mathbf{A}\mathbf{U}\Sigma\mathbf{V}^T\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{A}^T \\ &\approx \mathbf{A}\mathbf{U}\mathbf{A}\mathbf{U}^T\mathbf{A}^T \end{aligned} \quad (7)$$

The diagonal matrix \mathbf{A} contains the eigenvalues of $\Theta\Theta^T$ in non-increasing order. The eigenvectors in \mathbf{U} hold the importance of every tag with respect to a certain direction. Since $\Theta\Theta^T$ is symmetric, eigenvectors are chosen orthogonal to each other. Latent factors are thus incorporated into the tag information space by stretching it along the eigenvector directions according to the magnitude of the corresponding eigenvalues.

4. Application Possibilities

In this section, we describe several ways *TagMF* can be applied so that users may take benefit of tags in RS that rely on common model-based CF. The integrated model of latent factors and additional content information derived using our proposed method gives us the opportunity to access the previously abstract user-factor and item-factor vectors in a much more comprehensible manner: User profiles and item descriptions now comprise information related to both latent factors and user-generated tags. Thus, as the tag concept is easily understood by users, we can exploit the enriched vectors for several purposes: Among others, users may actively adjust their own user vector, i.e. indirectly determine their position in the latent factor space, in an interactive manner by means of tags according to their current situation. More concretely, in relation to the research questions formulated in Section 1, we enable users to ...

- select a small number of tags to express preferences at cold-start instead of rating items up front (RQ1a),
- weight tags to manipulate recommendations generated based on their existing user profile (RQ1b),
- critique a recommended item to receive suggestions that also take their profile into account (RQ1c),
- examine their preference profile by means of tag-based explanations (RQ1d).

In the following, we describe in detail how *TagMF* can be applied to realize interactive RS that support users in the different cases.

4.1. Eliciting Preferences at Cold-Start

In cold-start situations, users typically have to rate a certain number of items before CF recommenders can reliably predict their interests [38, 55] (see Section 2). When employing *TagMF*, new users can, in contrast to a conventional preference elicitation phase, be asked to select a (small) number of preferred tags to establish a user profile.

For this, we initialize a new user-tag vector a_u for a user u entering the system as follows:

$$a_{ut} = \begin{cases} 1 & \text{if tag } t \text{ has been selected by user } u \\ 0 & \text{else} \end{cases} \quad (8)$$

By multiplying this vector a_u with $\mathbf{U}\mathbf{\Lambda}^{1/2}$ (see (7)) holding the tag-factor relations, we obtain a regular user-factor vector. Now, to generate recommendations, this vector $a_u\mathbf{U}\mathbf{\Lambda}^{1/2}$ can be used the same way as if the vector p_u representing the user profile in standard MF had been derived exclusively based on ratings. This means we calculate its inner product with the item-factor vectors as shown in the introductory description in Section 3 (see also [24, 25]).

4.2. Manipulating Recommendations

For a user u with an existing preference profile based on explicit ratings or implicit behavioral data, i.e. a vector

p_u is already available, usually the only means to influence the recommendations in model-based CF systems is to (re-)rate single items (see Section 2). However, when a_u is derived by *TagMF* in the learning phase as described in Section 3.2, the user can additionally manipulate the entire result set in an interactive manner by means of tags provided by the community of users. This may support users in obtaining alternative suggestions, for instance, in case their long-term profile differs from actual interests or the recommendation list lacks diversity and novelty.

To this end, we define a weight vector $w_u \in [0, 1]^{|T|}$ that is supposed to capture user feedback in form of weights for tags, where 0 means no and 1 maximal interest of user u in tag t . For instance, a user who in the current situation is interested in action-packed movies that moreover contain a little more black humor than the ones usually recommended to him or her, may set the weights of the tags “action” and “black humor” to 1 and 0.5, respectively. This vector w_u can be then added to a_u in order to calculate recommendations based upon this update to the existing user profile. Consequently, we extend the original formulation (see again Section 3 as well as [24, 25]) as follows:

$$\tilde{r}_{ui} = (a_u + \pi w_u)\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T a_i, \quad (9)$$

with $\pi \in \mathbb{R}$ representing the degree to which the weight information is considered: We adaptively set π to 0 when the current user has not applied any weights, otherwise to $\frac{\|a_u\|}{n} \cdot \|w_u\|$, with $n \leq |T|$ being the number of tags already weighted by the user. Thus, when he or she sets all n weights to the maximum value, they have the same influence as a_u itself, i.e. both vectors are of equal length.

Provided users have any means to manipulate the values of w_u , e.g. sliders or spinners, the set of recommendations initially generated based on their long-term profile can now be continuously adapted in realtime, allowing them to interactively explore the effects of their preference settings, and, among others, to escape a potential “filter bubble”. Thereby, we in principle do not longer predict actual ratings: Only at the beginning, when all values of w_u are set to 0, \tilde{r}_{ui} effectively approximates r_{ui} . Instead, we combine the user’s general preference structure with the operationalization of his or her current interests or situational (e.g. mood- or activity-based) needs w_u , that he or she has expressed with respect to the tags by interacting with the system.

4.3. Critiquing a Recommended Item

Employing *TagMF* makes it further possible for users to interact with model-based CF systems in a more discrete fashion, resembling the well-known critiquing approach [21] (see Section 2.1). As in *MovieTuner* [18], an interactive variant based on user-generated tags implemented as part of the MovieLens³ platform, we are able

³<https://movielens.org/>

to let users request items that are overall similar to a currently recommended item i , but represent some selected dimensions less, equally or more strongly. This way, specific context-dependent or situational aspects of the search and decision process can be taken into account. For instance, in case the movie “Apocalypse Now” is shown, a user might apply the tag-based critique “less dark”, leading to “Saving Private Ryan” being suggested.

However, since our method builds on MF, we can additionally exploit the current user’s long-term profile. As a consequence, results presented after critiquing are not only related to the critiqued item (generally similar, but different with respect to applied critiques), but take to some degree this user’s general interests inferred from past user-item interaction data into account as it is customary for CF recommenders. Thus, considering the example from above, it might be that a user who tends to enjoy comedy more than other genres is presented with e.g. the movie “M*A*S*H” instead of “Saving Private Ryan” as a new recommendation. Moreover, the latent information available when using *TagMF* may influence the critiquing process in a way that resulting recommendations also reflect more subtle item characteristics that cannot be taken into account solely relying on explicit tag data.

Eventually, on condition that meaningful tags are somehow selected and presented as critique dimensions, it is necessary to reflect the critiques a user u has applied with respect to these tags to the currently recommended item i . For the implementation of this interaction mechanism, we combine the item-tag vector a_i with the user-tag vector a_u to a new vector a_c by performing the following steps⁴:

1. We scale a_i to the length of a_u , yielding a'_i . This ensures that in the end, we can still use a_c on the user side for generating recommendations.
2. Assuming that u likes the current suggestion due to very specific characteristics of i , we keep only values of a'_i that are two standard deviations above the mean of a'_i . All other entries are set to 0. Thereby, we avoid too homogeneous entries in a_c as it might be the case when just directly averaging all values of a'_i and a_u to combine them, which would lead to results neither related to i ’s characteristics nor u ’s profile.
3. We use a weighted average to combine a'_i with a_u , integrating a'_i with higher weight (here 60%) in order to more strongly reflect i ’s similarity to the items in the new result set. As the critiquing process continues, the weights may be dynamically adjusted.

Now, to generate recommendations, the resulting vector a_c can be used the same way as a_u before (see previous subsections). These recommendations are simultaneously geared towards i ’s characteristics as well as u ’s general interests regarding the tags. To also fulfill the user’s critique he or she has interactively applied, we employ the *linear-sat* variant of the critique distance (i.e. the difference of i

along the selected critique dimensions to the other items) as proposed in [18].

4.4. Explaining a User Profile

In systems relying on MF, users typically express their preferences indirectly with respect to the nontransparent latent factor space, e.g. through ratings for single items. The result are abstract user-factor vectors, making it difficult to explain a user’s profile. This can also be considered a common and more general issue in model-based CF. Our method, in contrast, allows to provide users with explicit tag-based explanations of the typically opaque representation of their long-term preferences within the model: As a consequence of taking additional content information into account, we can automatically determine those tags that are most important to an individual user—even if he or she never tagged any items.

For this purpose, we exploit that with *TagMF*, user-factor vectors are related to both latent knowledge and user-generated content, and thus become much more meaningful. Concretely, we utilize the matrix \mathbf{UA} holding the user-tag relations in order to explain the user representation as learned from historical user-item interaction data in textual form. When \mathbf{UA} is derived according to our method, this is independent of the tags a specific user actually has assigned: As described in Section 3.2, we derive tag-factor relations for all users by first learning the relationship between tags and latent factors, and then applying it to the user side. Hence, we can identify the most important tags for each user, even in the common case where he or she has not provided any tags but only conventional feedback (e.g. ratings). Thus, for the current user u , we select the n tags scoring highest in the corresponding user-tag vector a_u , and present them as a description of his or her long-term interest profile.

5. Evaluation

In order to answer the research questions posed in Section 1, we extensively evaluated *TagMF* both in offline experiments and empirical user studies.

First, to provide a basis for addressing RQ1, we performed offline experiments comprising an analysis of objective performance as well as a qualitative inspection of a resulting latent factor model. These experiments were supposed to show validity and general effectiveness of our method for enhancing model-based CF with additional content information.

Next, to analyze our method’s actual impact on users and to investigate the extended interaction possibilities provided, we conducted two empirical user studies: In the first study, we focused on the usage of *TagMF* for eliciting preferences in cold-start situations (RQ1a) and interactively manipulating recommendations that result from an existing user profile (RQ1b). In the second study, we investigated how *TagMF* can be applied to integrate model-based CF with critiquing, taking the recommended item

⁴We decided for the reported configuration due to pretests.

as well as the current user’s long-term interests into account (RQ1c). For these quantitative studies, we built an interactive web-based prototype movie RS that implements *TagMF*. To specifically examine the influence on subjective system aspects and user experience, we then performed a comparison with an automated recommender based on standard MF using ratings (RQ2a) in the first study, and with a tag-based interactive approach similar to *MovieTuner* (RQ2b) in the second one.

In the following, we describe all parts of this three-fold evaluation, concluding each with a detailed discussion that addresses the respective research questions.

5.1. Offline Experiments

Earlier experiments by others (see Section 2.2 and [e.g. 27, 29, 30, 35, 32, 34]) suggested that considering additional information improves accuracy of model-based CF recommendations as measured by common offline evaluation metrics. To confirm these findings and to validate our method’s effectiveness, we as well analyzed the objective performance⁵.

Moreover, while latent factors are generally considered to represent real-world characteristics [24, 26], we conducted a qualitative inspection of a factor model derived by means of *TagMF* to investigate whether automatically learning tag-factor relations according to our method actually leads to comprehensible and meaningful results.

5.1.1. Setup

In order to perform the experiments we used a Stochastic Gradient Descent MF algorithm⁶ based on [68] as a baseline. We extended this implementation of a common SVD-like MF algorithm according to our method as described in Section 3. As datasource for items as well as associated ratings and user-generated tags, we used the well-known MovieLens 20M dataset for ratings and the MovieLens Tag Genome dataset for item-tag relevance scores⁷. We then created an intersection of these datasets reducing them to items included in both, leaving us with 10 370 movies, 19 800 443 ratings and 11 697 360 tag relevance scores.

To run the performance analysis, i.e. to objectively compare standard SVD-like MF with *TagMF* in terms of recommendation accuracy, we used the RiVal benchmarking toolkit⁸ introduced in [69]. With this toolkit, we computed *Root Mean Square Error* (RMSE) [4] and *Normal-*

ized Discounted Cumulative Gain (NDCG) [4], a popular ranking metric from Information Retrieval.

5.1.2. Analysis of Objective Performance

First, we examined the influence of different basic configurations on objective recommendation accuracy using 10 % subsamples of users and 5-fold cross validation. We trained the standard MF and the *TagMF* models with 20 factors. For *TagMF*, we considered a limited number of the 50 most popular user-generated tags from the underlying dataset as additional training data. Figure 1 shows the experimental results for a comparison of standard MF and *TagMF* in terms of RMSE and NDCG@10 which we calculated as described above, varying the number of iterations and the regularization parameter λ when training the respective model.

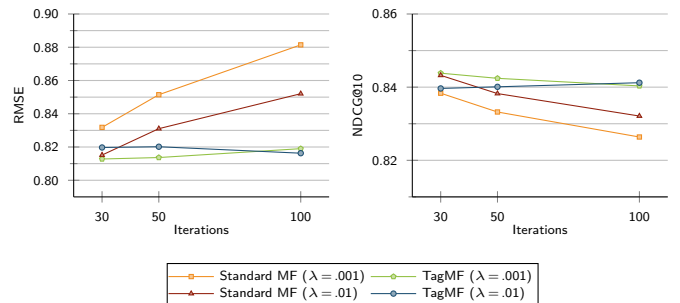


Figure 1: Comparison of standard MF and *TagMF* in terms of RMSE and NDCG@10 for different number of iterations and settings for λ .

Looking at these results, it seems that enhancing MF with additional information according to our method is beneficial. *TagMF* yields overall superior results both in terms of RMSE and NDCG@10. Furthermore, the results obtained with *TagMF* are rather stable. In contrast, iterating more often over the training data leads to decreased performance for standard MF.

Second, with 1 % subsamples of users, we performed another comparison of standard MF with *TagMF*, now varying the number of latent factors and the number of tags additionally considered in *TagMF*. Following further pretests, we used 30 iterations and set $\lambda = .03$. Then, we again performed 5-fold cross validation, yielding the RMSE and NDCG@10 results reported in Figure 2.

Overall, it again becomes apparent that considering additional information improves objective accuracy of MF. When using 50 tags or more, RMSE is lower for *TagMF* independent of the number of latent factors. NDCG@10 shows similar behavior, yielding equally promising results.

5.1.3. Qualitative Inspection

Enhancing a model-based CF recommender with additional content information according to our method may also help to gain a better understanding of the latent factor space⁹. Applying eigendecomposition as described in

⁵First results of offline experiments have been shown in the poster publication of [65]. Now, we present additional and more extensive experiments, among others with a newer and larger dataset.

⁶*ParallelSGDFactorizer* from the Apache Mahout recommender library (<http://mahout.apache.org/>).

⁷The MovieLens 20M dataset contains about 20 million ratings from more than 138 000 users for over 27 000 movies; The MovieLens Tag Genome dataset contains item-tag relevance scores for over 10 000 movies and 1 100 user-generated tags (<https://grouplens.org/datasets/>).

⁸<http://rival.recommenders.net/>

⁹We have briefly discussed this in the poster publication of [26].

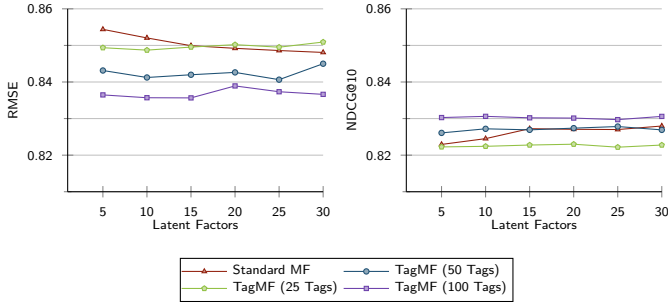


Figure 2: Comparison of standard MF and *TagMF* in terms of RMSE and NDCG@10 for different number of latent factors and tags.

Section 3.2 yields information on the importance of each dimension of the factor space and its relationship to the tags. Consequently, by examining the most positively and negatively related tags, respectively, we can obtain a more general understanding of what is expressed by factors derived automatically by means of MF.

Table 1 illustrates an example for relationships learned between factors and tags, resulting from a *TagMF* model trained on the MovieLens 20M dataset with 20 factors and 20 user-generated tags: Positive and negative values describe strength and direction of these relations. They express how strongly certain characteristics are represented within the respective factors, thus denoting their individual meaning. Apparently, the underlying semantics can be easily interpreted: For instance, while both factor 4 and 5 express characteristics related to “fantasy”, factor 4 has a very negative and factor 5 a very positive relation to the tag “action”. Accordingly, these two factors correspond to very different kinds of fantasy movies. This observation can be underpinned by extracting representative items for the respective factors, i.e. which have highest values in the item-factor matrix $\mathbf{AUA}^{1/2}$ (see Section 3.2). Here, for example, “Wizard of Oz” (factor 4) and “Star Wars: Episode IV – A New Hope” (factor 5) are clearly in line with the observed semantics.

5.1.4. Discussion

The offline evaluation generally shows that enhancing MF with additional information seems indeed beneficial in terms of objective recommendation quality. This is consistent with earlier retrospective offline experiments (see Section 2.2), validating the work of others [e.g. 27, 29, 30, 35, 32, 34] and providing a basis to further investigate RQ1.

With the rather limited subsamples of rating data used in our analysis, the decreasing accuracy of standard MF in Figure 1 compared to the largely stable results of *TagMF* is likely attributed to overfitting: Additional tag-based information appears to contribute more to control overfitting than increasing λ for standard MF. Also, as can be seen in Figure 2, the number of latent factors clearly has an influence on the performance of standard MF: The results improve with more factors and become stable only with 15

to 20 factors, while this parameter does not seem to affect *TagMF* to a large degree. With the amount of training data used, the number of tags incorporated according to our method seems to be the predominant factor for model quality. Nevertheless, with few tags (25 and 50), RMSE for *TagMF* goes up slightly when increasing the number of factors. Apparently, the variance in the factors cannot be covered sufficiently by the tags when there are fewer tags than factors. Thus, more factors appear to require considerably more tags to ensure consistently high model quality. Accordingly, in the example in Table 1, each factor is strongly related to multiple tags. However, parameter tuning in general, and e.g. determining an optimal ratio of factors to tags, is subject to future work. Overall, while observed differences between standard MF and *TagMF* are rather small independent of the number of tags taken into account (see again Figure 2), one can expect them to significantly increase when using a larger set of ratings. Then, as the data-generating function gets more complex, including more factor dimensions can be assumed to gain impact.

The qualitative inspection of the integrated *TagMF* model suggests that additional information may actually contribute to opening up the “black box” such models usually constitute for users in CF systems. The derived relations seem useful for the purpose of explaining latent factors through easily comprehensible user-generated tags. Moreover, the regression-constrained formulation (see Section 3.1) allows to gain insights on how users and items are positioned inside the latent space.

As shown in Table 1, we found items to be representative for the different dimensions and their relationship to the tags. With *TagMF*, latent factors are related to the tag information space by eigenvectors and eigenvalues (see Section 3.2), making it possible to translate positions for users the same way as for items. Thus, the method we use to derive user-tag relations (see again Section 3.2) ensures that users are assigned to equally meaningful positions. Accordingly, we proposed in Section 4.4 how to exploit these user-tag relations to select tags that explain a user’s long-term preference profile, thereby addressing the corresponding research question (RQ1d). While our approach consequently appears to be indeed a promising means to present users with explicit tag-based descriptions of their—in model-based CF typically nontransparent—preference profile, further validating this application possibility seems necessary.

5.2. Empirical User Study I

We performed the first user study to examine the influence considering additional information has on users in model-based CF systems, and to evaluate the interactive features that become possible by using *TagMF* in comparison to a conventional recommendation process¹⁰.

¹⁰This study has in large parts been presented in [70]. Now, we present more results and additional insights.

Table 1: Example for automatically learned relationships between latent factors (rows) and user-generated tags (columns): The five most important factors are shown together with positively (green) and negatively (red) related tags, as indicated by \mathbf{U} . The factor importance (in brackets in the left-most column) is equal to the values in $\mathbf{\Lambda}^{1/2}$. Representatives for each factor are automatically determined by extracting the movies (with at least 10000 ratings) that score highest for the respective factor in the actual item-factor matrix $\mathbf{I}\mathbf{A}\mathbf{U}\mathbf{\Lambda}^{1/2}$.

Factor	action	atmospheric	based on a book	classic	comedy	dark comedy	disturbing	dystopia	fantasy	funny	psychology	quirky	romance	sci-fi	surreal	time travel	thought-provoking	twist endings	violence	visually appealing	Automatically extracted representatives
1 (1.66)	0.25	0.38	-0.14	0.47	-0.20	0.16	0.14	0.04	-0.27	-0.15	-0.09	0.17	-0.26	-0.03	0.15	-0.19	0.06	-0.36	0.11	0.24	The Shining, Taxi Driver, A Clockwork Orange
2 (1.51)	-0.11	-0.12	0.02	-0.34	0.12	0.21	0.26	-0.12	0.27	-0.13	-0.02	0.14	-0.30	0.22	0.36	-0.51	-0.06	0.09	0.14	-0.21	Natural Born Killers, Brazil, Beetlejuice
3 (1.30)	0.10	0.11	-0.13	-0.63	-0.10	0.11	-0.06	0.07	-0.16	0.06	-0.07	0.21	-0.03	-0.16	0.18	0.17	-0.11	-0.05	-0.07	0.59	Amélie, Sin City, Magnolia
4 (1.21)	-0.39	-0.06	0.24	0.12	-0.16	0.00	0.03	-0.12	0.50	-0.22	0.05	0.14	0.29	-0.17	0.17	0.02	-0.08	-0.04	-0.48	0.19	Wizard of Oz, Willy Wonka & the Chocolate Factory, The NeverEnding Story
5 (1.17)	0.44	0.17	0.01	0.13	-0.11	-0.29	-0.16	0.01	0.44	0.10	-0.12	-0.27	-0.42	0.15	0.02	-0.16	0.01	-0.05	-0.20	0.28	Star Wars: Episode IV – A New Hope, Hobbit: An Unexpected Journey, Thor: The Dark World

5.2.1. Goals

First, we laid our focus on validating application possibilities of *TagMF*: For examining the value of user-generated tags as a means to elicit preferences at cold-start (RQ1a) and to interactively manipulate recommendations based on an existing user profile (RQ1b), we implemented them according to Section 4.1 and 4.2, respectively, in our a web-based prototype movie RS. Next, since we were interested in comparing the impact of additional information on subjective system aspects and resulting user experience to an automated rating-based CF recommender as it is common today (RQ2a), we formulated the following hypotheses contrasting this baseline and *TagMF*:

- H1:** *TagMF* improves perceived quality of recommendations.
- H2:** *TagMF* improves satisfaction with the item chosen from the recommendations.
- H3:** *TagMF* decreases difficulty to choose an item.
- H4:** *TagMF* has no negative impact on perceived interaction effort.
- H5:** *TagMF* improves transparency, especially in cold-start situations.

5.2.2. Method

The study was designed as an experiment under controlled conditions. We recruited 46 participants (33 female) with an average age of 22.89 ($SD = 6.88$), most of them students (85%). To interact with the prototype RS and to answer questionnaire items, participants used a common web browser at a desktop PC with a 24" LCD (1920 × 1200 px resolution). In the following, we describe the prototype system, the procedure, and the questionnaire we used, in more detail.

Prototype. Figure 3 shows the web-based prototype movie RS we implemented for the first user study. We set up two variants: One with a standard SVD-like MF algorithm [24], allowing users only to rate items. The interface resem-

bled a typical automated recommender based on ratings, with no interface elements related to tags present. This variant served as a baseline to test our hypotheses. The other variant was implemented based on *TagMF*. In order to validate the application possibilities, we extended this variant in comparison to contemporary model-based CF systems with several tag-based interaction mechanisms, as described in Section 4.1 and 4.2.

Concretely, the interface of the prototype is structured as follows: At the top (a), an area is shown where—in the *TagMF* variant—users can place tags and subsequently adjust their weight by means of sliders attached to them, this way manipulating the values of w_u (see Section 4.2). Note that in our prototype, it is not possible for users to create tags themselves, but only to use tags from the underlying dataset of tags provided by other users (see below). As *TagMF* can easily be applied with any set of tags, including tags generated by users of the respective system, this would indeed be different in a real-world scenario. Below (b), an input field allows to manually search for tags other users have applied, supported by autocompletion. These tags may be chosen to be weighted, i.e. to be placed in the area at the top together with a slider. In addition, the system initially suggests the 7 most popular tags, i.e. that have been assigned most often by users. As soon as the current user weights some tags, tags similar in terms of item-tag relevance data are suggested. The dialog in the top-right corner (c) presents users with a tag cloud describing their existing preference profile by means of tags chosen as described in Section 4.4.

Beneath, independent of the variant, the top-10 recommended items (d) are displayed together with movie poster and metadata. To further refine their profile, users may rate recommended movies and explicitly search further titles in order to rate them as well. In the *TagMF* variant, alongside each recommendation, the 3 most relevant tags for the respective movie are additionally shown (which may also be chosen to be weighted). Each manip-

ulation updates the result set immediately, thus providing users with direct and meaningful feedback regarding the effects of their preference settings on the recommender.

For calculating recommendations based on ratings, we used the same baseline algorithm as in the offline evaluation (Section 5.1.1). For the *TagMF* variant, we extended this algorithm according to Section 3, and implemented the interactive features as described in Section 4. Pretests similar to the offline experiments presented in Section 5.1, but based on the MovieLens 10M dataset¹¹, suggested to use 20 factors, 40 iterations, $\lambda = .001$, and to consider the 25 most popular user-generated tags from the underlying dataset as additional training data. We used the MovieLens 10M dataset for ratings and the MovieLens Tag Genome dataset for associated tags¹². We created an intersection of these datasets reducing them to those movies included in both, leaving us with 8 429 items, 9 964 745 ratings and 9 507 912 tag relevance scores. For the purpose of the study, we used scores precomputed as described in [71] based on user-generated tags from the underlying dataset. In a real-world scenario, one would indeed calculate the scores based on tags provided by the user community of the respective system, and then apply *TagMF* accordingly.

Questionnaire and Log Data. The questionnaire participants were required to fill in was primarily based on the pragmatic evaluation procedure for RS described in [72], containing items related to subjective system aspects and user experience. This evaluation framework (see Section 2.4) is based on [17], but is reduced to stable operationalizations of the subjective constructs and appears (after repeatedly being validated) to measure user experience in RS reasonably well with a limited number of questionnaire items [72]. Concretely, we assessed *Perceived Recommendation Quality*, *Choice Satisfaction*, *Choice Difficulty* and *Effort* by means of items from this framework. We used an additional item from [73] to assess recommendation *Transparency*.

We generated items ourselves to explicitly ask which of the two variants of the prototype RS participants prefer in general, and to let them rate the suitability for different situations of use (with or without search goal). To specifically analyze the usability of the additional interaction mechanisms, we applied the *System Usability Scale* (SUS [74]) and the *User Experience Questionnaire* (UEQ [75]) for the *TagMF* variant. In addition, we used again items from [73] to assess interface adequacy. Besides, we gathered data about demographics, interest in movies, and familiarity with the movie domain. Apart from UEQ (7-point bipolar scale), all items were assessed on a positive

5-point Likert-scale (1–5). We also collected qualitative feedback: An open-ended question asked participants to report suggestions and complaints. We logged user interaction behavior and measured task times.

Procedure. First, each participant was asked to complete two preliminary tasks in counter-balanced order that served to elicit an initial set of preferences both in form of numerical ratings (like in other CF systems) and preferred tags:

- a) Participants had to rate 10 out of the 30 most popular movies, which is a common value for a number of ratings that already leads to appropriate results [38]. We used these ratings for online-updating the factor vectors as proposed in [49]. Items were shown in random order and could be skipped when unknown.
- b) Participants had to select 3 tags¹³ they liked out of the 20 most popular ones from the dataset (shown in random order), which we used to initialize a meaningful user-tag vector a_u as described in Section 4.1.

Next, based on the two system variants implementing a standard MF algorithm and *TagMF*, respectively, we assigned participants in counter-balanced order to three different conditions in a within-subject design:

Standard MF: Standard SVD-like MF with initial recommendations based on the 10 ratings. The only interaction possible was to rate more items.

TagMF-Ratings: Tag-enhanced MF with initial recommendations based on the 10 ratings. Participants could again rate more items, but in addition weight tags in an interactive manner.

TagMF-Tags: Tag-enhanced MF with initial recommendations based on the 3 selected tags. Interaction mechanisms were equivalent to the previous condition.

In each condition, participants were initially presented with the top-6 recommendations generated by means of the respective algorithm. First, they were asked to choose one movie from these suggestions they would actually like to watch. Second, they rated their satisfaction with each of the movies on a 5-point Likert-scale (1–5). Third, they filled in the questionnaire described above regarding their subjective assessment of system and recommendations.

Next, in the interaction phase, participants were presented with the interface of the prototype variant that corresponds to the respective condition, showing the top-10 recommended movies (see Figure 3). Their task was to interact with the system using the provided means in order to further refine the recommendations and to receive a result set that better matched their personal interests. Eventually, participants finished the interaction phase at their own discretion.

¹¹At the time we conducted the first user study, not all data was yet released for the MovieLens 20M dataset.

¹²The MovieLens 10M dataset contains about 10 million ratings from more than 70 000 users for over 10 000 movies; The MovieLens Tag Genome dataset contains item-tag relevance scores for over 10 000 movies and 1 100 user-generated tags (<http://grouplens.org/datasets/>).

¹³For the number of tags to be selected, we analyzed the general interest of all users in the dataset regarding tags stored in $\mathbf{U}\mathbf{A}$ derived according to Section 3.2, and determined the tags with highest influence. We assume that such characteristic tags have a value at least one standard deviation above the mean of $\mathbf{U}\mathbf{A}$, leaving us with 3.46 tags per user.

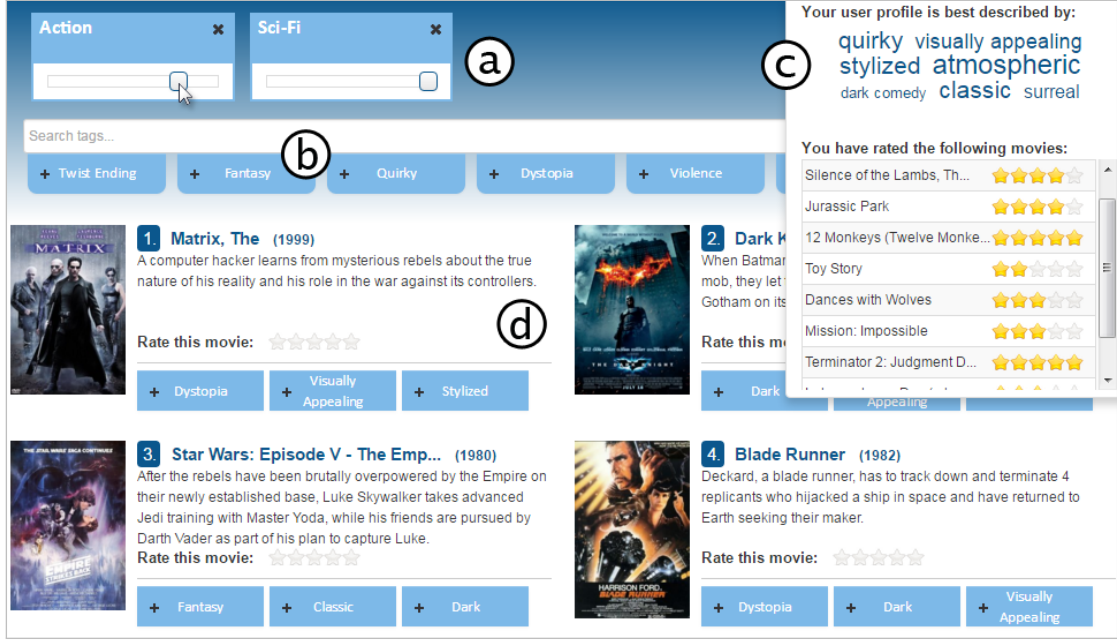


Figure 3: Screenshot of the prototype RS for the first user study: The current user has weighted the tags “action” and “sci-fi” (a), therefore receiving matching movie recommendations such as “Matrix” or “Star Wars” (d). The user can also search for other tags provided by the user community or get inspiration from the suggestions (b). Furthermore, the user’s existing profile is explained by a tag cloud (c).

Afterwards, participants were presented with the now adjusted top-6 recommendations. Again, they had to settle on one movie, rate how satisfying each recommendation was, and fill in a questionnaire¹⁴. Note that this time, the questionnaire was complemented with items regarding the interaction process.

5.2.3. Results

Participants reported that they liked movies a lot ($M=4.22$, $SD=0.63$) while having average knowledge about movies in general ($M=3.07$, $SD=0.80$) and about newer movies ($M=2.93$, $SD=0.98$).

We conducted two-way repeated measures ANOVA to compare the effects of condition and point in time on specific dependent variables corresponding to our hypotheses. For the comparison between the three conditions, mean values and standard errors are reported in Table 2.

In the following, we elaborate on the statistical significance ($\alpha=.05$) of the differences found in these results. Moreover, we report differences with respect to point in time. Note that interaction terms between the two factors were never significant, so we omit presenting them. For post hoc comparisons, we used the Bonferroni test.

Perceived Recommendation Quality. Concerning the subjective assessment of recommendations, there was a statistically significant effect for condition, $F(2, 90) = 7.40$, $p < .001$, $\eta_p^2 = .14$, with large effect size. Post hoc tests

Table 2: Mean values and standard errors for the different conditions. Higher values indicate better results (*Choice Difficulty* and *Effort* are reversed accordingly), except for time values additionally reported.

Construct	Standard MF		TagMF-Ratings		TagMF-Tags	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Perc. Rec. Quality	3.16	0.11	3.31	0.13	3.65	0.10
Mean Item Rating	3.11	0.10	3.29	0.11	3.55	0.10
Choice Satisfaction	4.00	0.10	4.10	0.13	4.35	0.09
Choice Difficulty	3.19	0.15	3.03	0.15	3.30	0.15
	33.82 s	3.09	28.41 s	2.60	28.48 s	2.37
Effort	3.77	0.13	3.84	0.10	3.64	0.11
	165.54 s	16.9	224.96 s	20.05	194.41 s	19.21
Transparency	3.20	0.15	3.41	0.15	3.73	0.13

indicated that the mean value for *TagMF-Tags* was significantly higher than for both, *TagMF-Ratings*, $p=.028$, and standard MF, $p<.001$. This confirms H1.

There was no significant difference regarding point in time, i.e. between before and after the respective interaction phase, $F(1, 45) = 0.02$, $p=.904$, $\eta_p^2 = .01$.

Mean Item Rating. With respect to ratings participants provided for each of the top-6 recommended items, we found differences to be similarly significant, $F(2, 88) = 11.19$, $p < .001$, $\eta_p^2 = .20$, again with large effect size. Movies in the *TagMF-Tags* condition received significantly higher ratings than in the two other conditions, *TagMF-Ratings*, $p=.025$, and standard MF, $p<.001$. As a consequence, we can eventually fully accept H1.

We found no significant effect with respect to point in time, $F(1, 44) = 0.02$, $p=.885$, $\eta_p^2 = .01$.

¹⁴For each condition, the dependent variables were thus assessed at two different points in time, i.e. before and after the respective interaction phase.

Choice Satisfaction. Regarding satisfaction with the movie participants finally selected from the set of recommendations, we found statistical evidence for differences between conditions, $F(2, 90) = 4.72$, $p = .011$, $\eta_p^2 = .10$, with medium effect size. Post hoc tests indicated that the mean value for *TagMF*-Tags was significantly higher than for standard MF, $p = .009$, which confirms H2. No differences were found between *TagMF*-Ratings and other conditions.

Furthermore, we found a significant difference regarding point in time, $F(1, 45) = 5.07$, $p = .029$, $\eta_p^2 = .10$, with medium effect size. Before interaction phases ($M = 4.28$, $SE = 0.10$), participants were more satisfied with their selected movie than afterwards ($M = 4.02$, $SE = 0.11$).

Choice Difficulty. We objectively operationalized the difficulty to decide as the total time participants spent for settling on a movie they would actually like to watch from the shown top-6 recommendations. The within-subjects main effect yielded significant differences with medium effect size for condition, $F(2, 88) = 5.34$, $p = .006$, $\eta_p^2 = .11$. Participants took significantly more time with standard MF compared to *TagMF*-Ratings, $p = .015$, and *TagMF*-Tags, $p = .050$. The difference between the two *TagMF* conditions was not significant.

Participants decided more quickly after the interaction phases ($M = 25.81$ sec, $SE = 2.31$) than before ($M = 34.66$ sec, $SE = 2.88$), with significant difference and large effect size, $F(1, 44) = 28.03$, $p < .001$, $\eta_p^2 = .39$.

In addition, we specifically asked participants how difficult it was to choose a movie¹⁵: With respect to their subjective perception, we neither found a significant effect for condition, $F(2, 90) = 1.20$, $p = .307$, $\eta_p^2 = .03$, nor point in time, $F(1, 45) = 1.60$, $p = .212$, $\eta_p^2 = .03$. Overall, we can thus only partly accept H3.

Effort. Concerning total time participants spent in the different conditions for the interaction phase, we found a significant effect using a one-way ANOVA, $F(2, 90) = 3.34$, $p = .040$, $\eta_p^2 = .07$, with medium effect size. On average, participants needed significantly more time in the *TagMF*-Ratings condition compared to standard MF, $p = .040$. No differences were found in other pairwise comparisons.

However, although the interaction phase in both *TagMF* conditions was at least marginally longer, we found no significant differences with respect to perceived interaction effort¹⁵, which we assessed after each interaction phase and analyzed using a one-way ANOVA, $F(2, 90) = 1.40$, $p = .253$, $\eta_p^2 = .03$. Overall, this confirms H4.

Transparency. Once again using a two-way repeated measures ANOVA, we noted a significant effect of condition on transparency, $F(2, 90) = 6.22$, $p = .003$, $\eta_p^2 = .12$, with medium to large effect size. Results from standard MF were perceived less transparent than from *TagMF*-Tags,

$p = .003$, which confirms H5. No differences were found between *TagMF*-Ratings and other conditions.

Moreover, no significant effect was found for point in time, $F(1, 45) = 0.01$, $p = .948$, $\eta_p^2 = .01$.

Usability. Regarding the two variants of our prototype RS, a paired t -test ($t(45) = 4.15$, $p < .001$) indicated that participants generally preferred the variant that integrated interactive features based on *TagMF* ($M = 3.76$, $SD = 1.02$) over the one that used standard MF ($M = 2.83$, $SD = 1.00$). Some participants, for instance, explicitly stated that they “do not want to use only star ratings, but rate several aspects, so that the system can better recommend movies” and “really liked the tag selection with the sliders”.

More specifically¹⁶, usability of the prototype variant that supported interaction via tags was rated as “good” with a SUS score of 78. Values between 0.95 and 1.96 on the different subscales of the UEQ were equally promising. In particular, the subscale for transparency yielded an “excellent” score ($M = 1.96$), and efficiency was rated “above average” ($M = 1.16$), which corresponds to the very positive assessment of interface adequacy ($M = 4.13$, $SD = 0.48$). Overall, the variant was rated to be particularly useful with no ($M = 3.78$, $SD = 0.99$) or only a vague ($M = 3.89$, $SD = 1.02$) search goal in mind. In contrast, but as expected, participants found it less suitable when they already knew their search direction ($M = 2.52$, $SD = 1.50$).

5.2.4. Structural Equation Modeling

Since we were particularly interested in differences between conditions in cold-start situations where the system must deal with high uncertainty, we used SEM (see Section 2.4) to further investigate the effects of using different recommender algorithms and methods for eliciting initial preferences on subjective system aspects and user experience.

Based on the framework for user-centric evaluation of RS proposed in [17] (see Section 2.4), we defined algorithm (*Standard MF vs. TagMF*) and initial preference elicitation method (*Ratings vs. Tags*) as *Objective System Aspects* (OSA) that cannot be influenced by the user. We considered *Perceived Recommendation Quality* and *Transparency* as *Subjective System Aspects* (SSA) representing user perception of OSA. SSA are seen as mediating variables between OSA and user experience [17, 16]. User experience is known to be substantially affected by underlying algorithms and preference elicitation methods [e.g. 76, 17, 21, 77, 63, 16, 78]. In light of this, we assumed user experience and interaction behavior to be influenced through changes regarding SSA, when using, for example, a novel means for eliciting initial preferences such as selecting tags according to Section 4.1. Consequently, we

¹⁵Note that higher values indicate better results.

¹⁶Note that we only asked specific questions regarding usability for the *TagMF* variant to reduce participants’ workload in the within-subject design. Besides, interaction in the other variant was limited to rating items, minimizing the need for a separate evaluation.

included *Choice Satisfaction* as an indicator of *User Experience* (EXP), and complemented the more general perceived quality of the set of top-6 recommendations by capturing *Interaction Behavior* (INT) in form of the specific rating feedback for the individual movies, i.e. *Mean Item Rating*. In addition, we took personal characteristics into account to deduce assumptions about the influence of different dispositions. In line with the underlying framework, we assumed attitude and behavior concerning the varied system aspects to be affected by certain *Personal Characteristics* (PC) such as *Domain Knowledge* and *Trust in Technology*.

We set up a first theoretical model (Figure 4), yielding a good fit with the data ($\chi^2(7) = 8.246$, $p = .311$, $CFI = .995$, $TLI = .989$, $RMSEA = .032$). It explains a large amount of variance regarding our dependent variables *Choice Satisfaction* ($R^2 = .408$) and *Mean Item Rating* ($R^2 = .698$), as well as about 20% of our considered mediator *Perceived Recommendation Quality* ($R^2 = .208$).

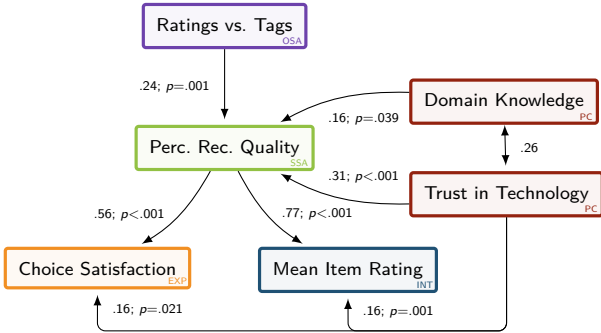


Figure 4: Path model for comparing the influence of initial preference elicitation via ratings or tags. On the edges, standardized regression weights as well as corresponding p -values are displayed.

Direct effects of varying the algorithm (*Standard MF vs. TagMF*) between conditions were not significant for any dependent variable or the mediator. Thus, this OSA was eventually not integrated in our model. The method for initial preference elicitation (*Ratings vs. Tags*) seems in contrast to account for a significant explanation of *Perceived Recommendation Quality*. Regarding personal characteristics, *Domain Knowledge* shows a meaningful influence only on *Perceived Recommendation Quality*, but *Trust in Technology* on all dependent variables.

The mediator *Perceived Recommendation Quality*, an overall subjective assessment, seems to be a strong predictor for both more specific variables, *Choice Satisfaction* and *Mean Item Rating*. Further analysis shows that *Perceived Recommendation Quality* appears to completely mediate the otherwise significant predictive power of varying the initial preference elicitation method (*Ratings vs. Tags*).

In view of our hypotheses, we aimed at further clarifying the role of participants' understanding of recommendations in cold-start situations (H5). Thus, we integrated *Transparency* as additional mediator in a second

model (Figure 5). Overall, this model, which again fits the data well ($\chi^2(12) = 13.669$, $p = .322$, $CFI = .995$, $TLI = .989$, $RMSEA = .032$), explains a large proportion of variance regarding *Choice Satisfaction* ($R^2 = .401$), *Mean Item Rating* ($R^2 = .693$) and *Perceived Recommendation Quality* ($R^2 = .523$). Moreover, it achieves a reasonable amount of explained variance with regard to *Transparency* ($R^2 = .234$).

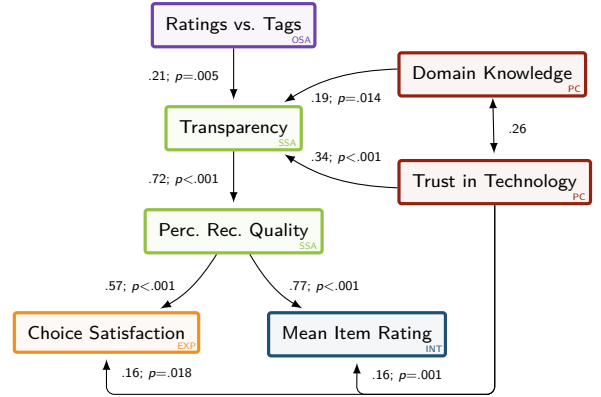


Figure 5: Path model for comparing the influence of initial preference elicitation via ratings or tags, mediated by transparency. On the edges, standardized regression weights and p -values are displayed.

The second proposed model shows that the predictive power of *Perceived Recommendation Quality* on the dependent variables observed in the first model obviously still holds. However, there are significant shifts of relations between the variables due to integrating *Transparency*: *Transparency* seems to be a substantial causal factor for *Perceived Recommendation Quality*, which in turn acts as a complete mediator for the effects on our more specific dependent variables. In fact, *Transparency* itself seems to be a regressor fully mediating the direct effect of varying the initial preference elicitation method (*Ratings vs. Tags*) on *Perceived Recommendation Quality* found in the first model. This confirms H5 even for the special case of cold-start. Besides, *Transparency* appears to be a partially mediating variable for the personal characteristics *Domain Knowledge* and *Trust in Technology*.

5.2.5. Discussion

In the first user study, the variant of our prototype system that relied on *TagMF* received significantly higher scores with respect to a number of variables related to subjective system aspects and user experience. Even in cases differences were found not significant between standard MF and *TagMF*-Ratings, results in the condition based on the extended variant tended to be better¹⁷. Regard-

¹⁷In all conditions, some scores were not as high as expected. We assume this to be due to the dataset (only movies released before 2008) and our particular sample of participants (more females, rather young, average domain knowledge). Qualitative answers to the open-ended question as well as better results in the second user study (newer dataset, more homogeneous sample) support this assumption.

ing assessment before and after interaction phases, perceived recommendation quality and transparency did not differ significantly. Since interaction terms of condition and point in time were never significant, we deduce that this applies to all conditions. Satisfaction with the chosen movie was even decreased after interaction phases¹⁸, but this can be justified by examining typical user behavior: Participants rated movies they knew and liked already during these phases. Consequently, the result set changed a lot, eventually comprising items which might be not as easy to assess at first sight. One participant explicitly mentioned that the recommendation set “would have better fitted his or her taste when movies he or she rated highly had not been removed”. Still, scores related to recommendation quality (H1), choice satisfaction (H2) and transparency (H5) seem very promising, and more importantly, higher in both *TagMF* conditions.

In real-world scenarios, initial preference elicitation—here performed as a preliminary task—would be part of actual system use. In this context, the significant differences between conditions before the interaction phase (with *TagMF* being superior) suggest that the few interaction steps initially taken, i.e. selecting a small number of tags up front, are already sufficient to improve user experience of typical RS. In particular, transparency in the *TagMF* conditions was rated better even at the beginning, although participants did not know that results were based exclusively on few tags. Thus, considering additional content information according to our method seems to help users implicitly when judging recommendations independent of later interaction (H5).

Because of these findings, we further examined the role of transparency at cold-start by using SEM. Our first proposed model indicated that selecting tags instead of rating items to elicit initial preferences significantly improves perceived recommendation quality (H1). Including transparency into the second model increased the amount of variance explained by the entire model concerning perceived recommendation quality from 21 % to 52 %. With a high standardized regression weight, transparency appears to be a substantial predictor for perceived recommendation quality. In turn, varying the preference elicitation method significantly contributes to explaining transparency (see Figure 5). The second model further shows that the effect on perceived recommendation quality found in the first model is actually fully mediated by transparency. Apparently, relying on *TagMF* leads to more comprehensible results (H5), which are consequently perceived to be of higher quality, ultimately also increasing participants’ satisfaction with their chosen item (H1, H2). We deduce that user-generated tags import semantics into the result set

¹⁸In terms of choice difficulty operationalized as the time spent for settling on a movie, the point in time also had a significant effect. However, this was expected as it is likely that participants already decided for an item during interaction phases, and were therefore able to choose faster when asked to choose a movie afterwards. Note that the subjective perception did not differ significantly.

which are more natural to understand than a meaning derived from recommendations calculated exclusively based on typical user-item interaction data. Our qualitative inspection of the factor space supports this (Section 5.1.3). In summary, the significant influence of initial preference elicitation method emphasizes that selecting tags according to the application possibility described in Section 4.1 instead of rating items up front is a promising means to alleviate the cold-start problem in model-based CF systems (RQ1a).

As a side note, while recommendation quality was indeed the main predictor for choice satisfaction and individual rating feedback, also personal characteristics had an impact. For instance, by increasing transparency, our method seems particularly useful for users with little domain knowledge, as it becomes easier to comprehend why certain items are recommended. The influence of trust in technology was however only partially mediated via transparency. Personal characteristics thus might alter the way perceived quality is translated into numerical ratings: Users whose trust in technology is low are likely to provide lower ratings in a more technically-oriented system. This poses another argument for using more natural ways to interact with CF systems than (re-)rating single items.

System usability was assessed overall very positively for the prototype variant based on *TagMF*¹⁶. Some participants had specific suggestions (e.g. “full text search should be integrated”) or complaints (e.g. “movies cannot be excluded from the results without rating them”) with regard to system functionalities. However, these qualitative comments addressed rather general usability issues beyond the scope of our research, and were, in particular, not exclusively related to the variant that supported interaction via tags. When asked specifically, participants in general preferred this variant. While this might be a reason why they spent more time in the two corresponding conditions (significantly or at least tendentially longer interaction phases), the richer interaction possibilities may account for this finding as well. Moreover, participants had to get used to the novel mechanisms introduced by *TagMF* while they were likely more familiar with conventional rating-based interfaces. Either way, perceived interaction effort did not differ significantly, so that we can accept H4.

In summary, our interactive approach realized by applying *TagMF* seems valuable for improving transparency of recommendations as well as providing users with extended possibilities to control the recommendation process and to adapt the results towards their current interests when relying on an existing long-term profile, thus validating the application possibility described in Section 4.2 (RQ1b).

Lastly, with our study, we for the first time confirmed that enhancing model-based CF with additional information is beneficial with respect to subjective perception of recommendation quality, which previously has only been observed in terms of offline performance (see Section 2.2).

For cold-start situations, SEM however showed no significant difference in this regard when varying the algorithm. This is generally in line with recent research stating that different or objectively more accurate recommenders do not necessarily produce better results from a user perspective [5, 6, 7, 63]. Although it may achieve high accuracy scores, a list of items detached from a superordinate context might not be satisfactory for users. Instead, the recommendation set should exhibit some kind of inner consistency, which in our case is reached through establishing a relationship between latent factors as derived by MF and user-generated tags. Recommendations thus seem to refer to each other implied by the easy-to-understand semantics of tags. Consequently, using *TagMF* positively affects transparency by building a meaningful context, thereby in turn improving perceived recommendation quality. In accordance with this, participants needed significantly less time to choose a movie from the recommendations in the respective conditions (H3). Beyond that, our study showed that compared to a typical automated RS based on ratings, also other subjective aspects related to user experience benefit equally from considering additional information according to our method (RQ2a).

5.3. Empirical User Study II

We performed the second user study with the goal of investigating the influence latent knowledge has on the recommendation process from a user perspective. In this regard, we wanted to focus on the comparison against an interactive RS that relies on user-generated content alone, and to examine the value of *TagMF* for implementing critiquing.

5.3.1. Goals

First, we aimed at validating another application possibility of *TagMF*: For evaluating the option to interactively critique a recommended item by means of user-generated tags in a model-based CF scenario (RQ1c), we implemented it according to Section 4.3 in our web-based prototype movie RS. Next, since we were interested in examining the impact using a latent factor model that integrates additional information has on the subjective assessment of system aspects, and thus on user experience, when compared to a purely tag-based interactive approach (RQ2b), we formulated the following hypotheses contrasting this baseline and *TagMF*:

- H1:** *TagMF* improves perceived quality of recommendations.
- H2:** *TagMF* improves satisfaction with the item chosen from the recommendations.
- H3:** *TagMF* decreases difficulty to choose an item.
- H4:** *TagMF* decreases perceived interaction effort.
- H5:** *TagMF* leads to more diverse recommendations.
- H6:** *TagMF* has no negative impact on transparency.
- H7:** *TagMF* improves perceived quality of critiquing.

5.3.2. Method

The study was designed as an experiment under controlled conditions. We had 54 participants (37 female) with an average age of 27.89 ($SD=10.30$), a small majority of them students (57%). To interact with the prototype RS and to answer questionnaire items, they used a common web browser running on a desktop PC with a 24" LCD (1920×1200 px resolution). Next, we describe the prototype system, the procedure, and the used questionnaire, in more detail.

Prototype. Figure 6 shows the web-based prototype movie RS we developed for the second user study. We again set up two variants: One reimplementing the method behind *MovieTuner* [18], with an interface resembling a typical critique-based RS (see Section 2.1), in particular, the integration of *MovieTuner* in the MovieLens platform [18]. This purely tag-based variant served as a baseline to test our hypotheses. The other variant with nearly identical interface was implemented using *TagMF*. Here, to validate this application possibility, we integrated the interactive critiquing process as described in Section 4.3.

Concretely, the interface is structured as follows: At the top (a), the critiquing area comprising tags used as dimensions to critique the currently recommended item is displayed. As in the *MovieTuner*, these tags generated by the user community are automatically shown by the system based on the method described in [18]. This method considers tag utility, popularity and diversity to determine a set of tags that is particularly meaningful for critiquing the current item. The only requirement is availability of item-related tag relevance scores, i.e. our given matrix \mathbf{A} (see Section 3.1). However, in the variant relying on *TagMF*, we additionally exploit that user-item interaction data is available as usual in CF systems, and blend this set together with a set of tags reflecting the current user's specific interests. Concretely, we replace half of the presented critique dimensions with tags scoring highest for this individual user, thereby personalizing the critiquing area. This is possible with *TagMF* as we also know user-related tag relevance scores for all available tags provided by the user community, i.e. our derived matrix \mathbf{A} (see Section 3.2). Either way, each tag is accompanied with radio buttons allowing users to critique the currently recommended item (details for this movie are shown on demand when hovering its title), i.e. requesting new suggestions with less, equal, or more relevance with respect to the corresponding tag (we implemented critiquing for the two system variants according to the respective method, as described below).

Moreover, users can search and manually choose tags as additional critique dimensions using the input field underneath (b), supported by autocompletion. As in the first user study (see Section 5.2.2), all available tags come from the underlying dataset (see below) and are generated by other users. Yet, since *TagMF* can be used with any set of tags, it would indeed be possible to also create new tags

in a real-world system.

In the *TagMF* variant of our prototype, the user profile is additionally presented in a dialog (c) similar as in the prototype for the first user study (see Section 5.2.2). The rest of the screen shows the top-9 recommendations. Note that in contrast to standard critique-based RS, and thus to the tag-based prototype variant, these recommendations in the *TagMF* variant rely on both the user’s situational needs, i.e. critiques applied to the currently recommended item, as well as his or her long-term profile based on historical preference data as it is customary in CF systems. Each recommendation (d) is displayed together with the 3 tags most relevant to the respective movie (these tags may be selected as critique dimensions as well), and a button that may be used to choose this movie as a new item to critique, i.e. to start a new cycle in the critiquing process.

For calculating recommendations in the *TagMF* variant, we again used a Stochastic Gradient Descent MF algorithm as point of departure. As a result of the offline experiments reported in Section 5.1, we used 20 factors, 30 iterations, and set $\lambda = .001$. Moreover, we used the 50 most popular user-generated tags from the underlying dataset as additional training data, and integrated the resulting model-based CF recommender with critiquing as described in Section 4.3. For generating recommendations and integrating the critiquing process in the other variant of our prototype system, i.e. the one exclusively based on tags, we reimplemented the method behind *MovieTuner* as proposed in [71, 18, 79]. For this, we again relied on the 50 most popular tags. According to prior testing, we chose the *linear-sat* metric for computing critique satisfaction. Further parameters are set as suggested in the literature. For item data, associated ratings and tag relevance scores, we used the same intersected dataset based on MovieLens 20M and MovieLens Tag Genome dataset as in the offline experiments (see Section 5.1.1). While we thus relied on scores precomputed as described in [71] based on user-generated tags from the underlying dataset, one would in a real-world scenario indeed use tags provided by users of the system at hand to calculate these scores.

Questionnaire and Log Data. As in the first user study, the questionnaire participants had to fill in was primarily based on the pragmatic evaluation procedure for RS proposed in [72], containing items related to subjective system aspects and user experience (see Section 5.2.2). Concretely, we assessed *Perceived Recommendation Quality*, *Choice Satisfaction*, *Choice Difficulty*, *Effort* and *Diversity* by means of items from this framework. We used an item from [73] to assess *Transparency* of recommendations. Regarding the *Critiquing*, we selected items from [18].

Again using items from [73], we assessed the overall satisfaction of participants with the respective prototype variant as well as the interface adequacy. In addition, we applied *System Usability Scale* (SUS [74]) and *User Experience Questionnaire* (UEQ [75]). We gathered data about demographics and familiarity of participants with

the movie domain. Apart from UEQ (7-point bipolar scale), all items were assessed on a positive 5-point Likert-scale (1–5). We also collected qualitative feedback: An open-ended question asked participants to report suggestions and complaints. Finally, we logged user interaction behavior and measured task times.

Procedure. First, participants were asked to complete a preliminary task that served to elicit an initial set of preferences. For this, movies were presented one after the other based on popularity and entropy as proposed in [80]. Items were separated into blocks of 25, and then shuffled to eliminate sequence effects. Unknown movies could be skipped. After participants rated 10 movies, this feedback was used to initialize a standard factor vector using online-updating (again implemented according to [49]) and to subsequently generate recommendations using *TagMF*: The top-15 results were presented in form of a list that could be expanded up to a maximum of 30 movies. Participants had to choose one movie out of these recommendations which they should know, and would find interesting as a starting point for a succeeding critiquing process.

Next, participants were assigned in counter-balanced order in a between-subject design to one of the two following conditions that correspond to the two system variants (yielding 27 participants per condition):

Tag-based: Tag-based method with recommendations based on similarity and critique distance to the currently recommended item in terms of tag relevance, implemented according to the method behind *MovieTuner* [71, 18, 79]. Critique dimensions were shown based on item-related tag relevance scores. Participants could interactively select tags, apply critiques, and switch the critiqued item.

TagMF: Tag-enhanced MF with recommendations based on user profile (i.e. derived factor vector) and currently recommended item as described in Section 4.3. Critique dimensions were suggested based on item-related as well as user-related tag relevance scores. Interaction was equivalent to the other condition.

In both conditions, participants were initially presented with an item representing the starting point for the critiquing process (see task descriptions below) as well as the top-9 recommendations generated according to the underlying method. Note that the only visible difference in the interface of the two prototype variants was availability of the dialog showing the user profile (see Figure 6). In the background, however, the way critique dimensions were selected and recommendations were generated differed as described above. Either way, participants had to interact with the respective system variant, i.e. apply critiques and switch the currently critiqued item, in order to refine the set of 9 recommendations and to fulfill the following tasks:

- 1) Participants were asked to find a movie that fits their personal preferences and they would actually like to watch. As a starting point, the movie chosen after the preliminary task was shown. Recommendations

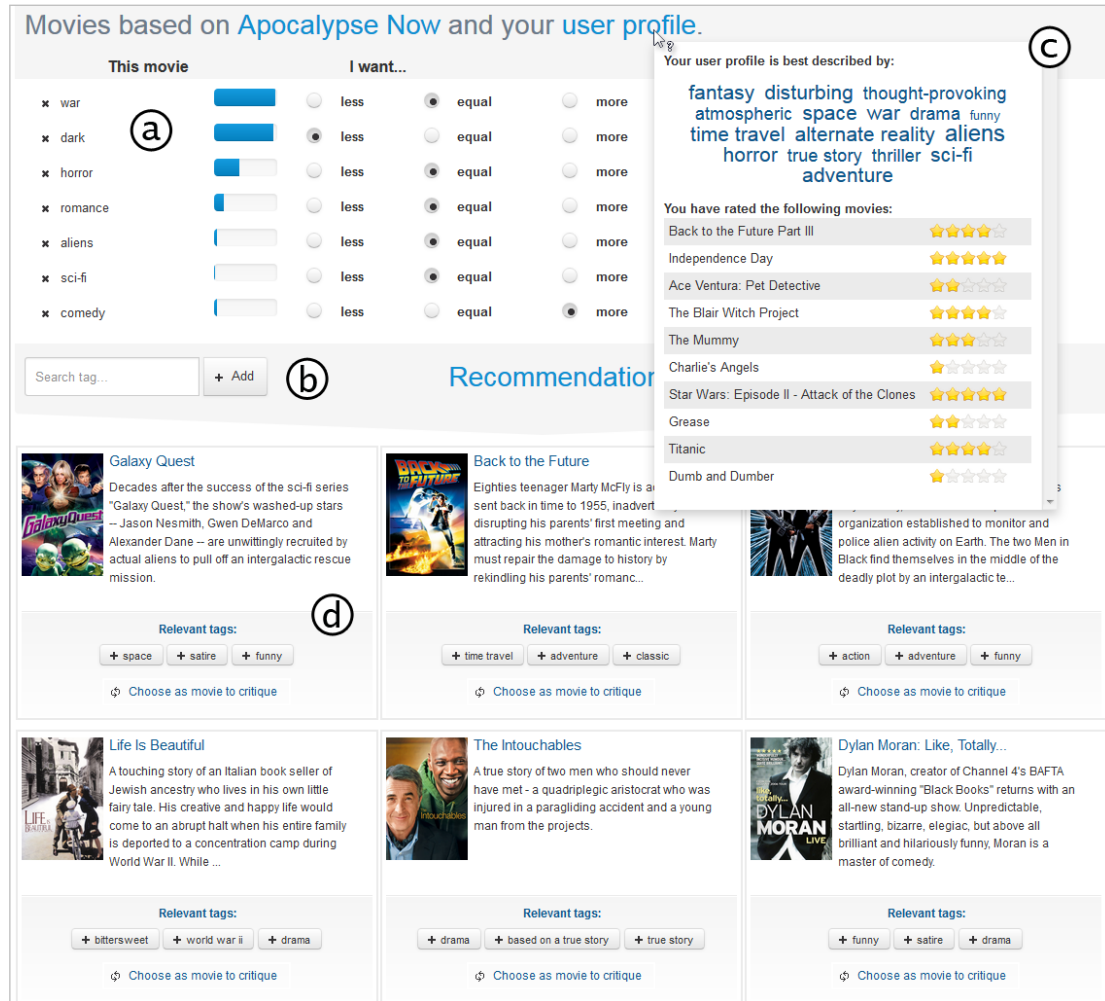


Figure 6: Screenshot of the prototype RS for the second user study: A user whose profile is shown in the dialog (c) has applied a critique (a) to the currently recommended movie “Apocalypse Now” using the tags “dark” and “comedy”. As a consequence, recommendations that fit to the critique and to his or her long-term interests are shown (d). To add further critique dimensions, the user can also search for tags provided by other users (b).

based on the currently critiqued item, and in the *TagMF* condition additionally on the preferences of the current participant, i.e. the factor vector learned by means of the 10 ratings elicited up front.

- 2a) Participants were asked to find a movie that they would like to watch when going out on a date with someone. Thus, they were required to not only take their personal preferences into account, but in addition interests of the fictitious date (which were not explicitly given). As a starting point, the movie chosen after the preliminary task was shown. Recommendations were generated as in the previous task.
- 2b) Participants were asked to find an adequate movie for the given situation that an adult horror movie fan wants to watch a movie together with a 9-year-old child. Thus, they were required to assume a high interest in horror movies while taking the interests of the child into account (which were not explicitly given). As a starting point, we selected a represen-

tative horror movie. Recommendations based on the currently critiqued item, and in the *TagMF* condition additionally on an artificial profile we created by training a factor vector with ratings typical for a horror movie enthusiast.

Task 2a and 2b were presented in random order. All tasks were finished by participants at their own discretion.

After each task, participants were first asked to choose the movie they found most suitable for the given task from the final set of top-9 recommendations. Second, they had to rate their satisfaction with each of the recommended items on a 5-point Likert-scale (1–5). Finally, participants were asked to fill in the questionnaire as described above.

5.3.3. Results

Participants of the second user study reported average knowledge about movies ($M=3.02$, $SD=0.87$). The movie chosen initially as a starting point after completing the preliminary task was rated very positively ($M=4.65$, $SD=0.68$), while most participants had seen it (94%).

For the directed hypotheses, we conducted one-tailed t -tests (using a significance level of $\alpha = .05$) to compare the two conditions in terms of corresponding dependent variables. In contrast to the first user study with a repeating task in a within-subject design, the different nature of the tasks in the second user study made a comparison between tasks rather meaningless. Instead, we were specifically interested in individual results per task. Thus, we omit reporting repeated-measures variance analyses, but present the results with respect to subjective system aspects and user experience separately for each task in Table 3.

Table 3: t -test results with mean values and standard deviations ($df = 52$, except for † (48.36) and ‡ (45.51) adjusted due to unequal variances) comparing the conditions with respect to subjective system aspects and user experience (* indicates significance at 5% level; d represents Cohen’s effect size value). Higher values indicate better results (*Choice Difficulty* and *Effort* are reversed accordingly).

Construct & Task	Tag-based		TagMF		T	p	d
	M	SD	M	SD			
Perc. Rec. Quality							
Task 1	3.67	0.84	4.20	0.67	2.59	.006*	0.71
Task 2a	3.87	0.93	4.19	0.86	1.30	.100	0.35
Task 2b	3.26	0.81	4.02	0.88	3.29	.001*	0.90
Mean Rating							
Task 1	3.61	0.55	3.83	0.66	1.32	.097	0.36
Task 2a	3.45	0.49	3.86	0.57	2.75	.004*	0.76
Task 2b	3.27	0.55	3.65	0.64	2.30	.013*	0.63
Choice Satisfaction							
Task 1	4.59	0.50	4.78	0.64	1.18	.121	0.32
Task 2a	4.56	0.64	4.81	0.48	1.68†	.050*	0.46
Task 2b	4.00	0.83	4.52	0.64	2.56	.013*	0.70
Choice Difficulty							
Task 1	3.59	1.01	3.22	1.28	-1.18	.122	-0.32
Task 2a	3.37	1.15	3.33	1.33	-0.11	.457	-0.03
Task 2b	2.89	1.09	3.19	1.30	0.91	.184	0.25
Effort							
Task 1	3.98	0.60	4.06	0.80	0.39	.351	0.11
Task 2a	3.89	0.87	4.09	0.75	0.92	.180	0.25
Task 2b	3.46	0.63	3.72	0.94	1.19‡	.121	0.32
Diversity							
Task 1	3.67	0.92	4.07	0.83	1.71	.047*	0.47
Task 2a	3.89	0.89	4.19	0.62	1.42	.082	0.39
Task 2b	3.81	0.79	4.11	0.75	1.42	.082	0.39

Perceived Recommendation Quality. Concerning perceived quality of recommendations, we found a statistically significant effect for condition in Task 1 and Task 2b. Mean values for *TagMF* were significantly higher than in the tag-based condition. Note that effect sizes were medium to large, or at least small to medium for Task 2a. Overall, this confirms H1.

Mean Item Rating. Individual ratings participants provided for the top-9 recommended items in the *TagMF* condition were found significantly higher in Task 2a and 2b, with medium to large effect sizes. Although there was no significant difference in Task 1, the ratings given in the *TagMF* condition were on average higher than in the tag-based condition, with small to medium effect size. Thus, we can eventually fully accept H1.

Choice Satisfaction. Participants in the *TagMF* condition were more satisfied with the movie chosen from the set of top-9 recommendations in all tasks. For Task 2a and 2b, we even found statistical evidence for differences between the tested conditions, with medium to large effect sizes. Overall, this confirms H2.

Choice Difficulty. Regarding the subjective assessment of the difficulty to choose one item from the set of movies eventually recommended¹⁵, we found no significant differences between conditions. In two comparisons, the tag-based variant received marginally better results, but with rather small effect sizes. Nevertheless, we have to reject H3.

Effort. Interaction effort was perceived slightly better in the *TagMF* condition¹⁵. However, we could not observe significant differences. This is reflected in task times, which likewise did not differ between conditions: Task 1: $t(52) = 1.24$, $p = .111$, $d = 0.34$; Task 2a: $t(52) = -0.25$, $p = .401$, $d = -0.07$; Task 2b: $t(52) = 1.06$, $p = .147$, $d = 0.29$. Overall, while *TagMF* at least tended to get better subjective results in all tasks, we have to reject H4.

Diversity. Participants rated the diversity of the set of recommendations generated by *TagMF* higher than participants in the tag-based condition. With medium effect sizes for all tasks, we even found a significant difference between the two conditions in Task 1. Overall, this confirms H5.

Transparency. Once after completing all tasks, we asked participants how they perceived the transparency of recommendations. They provided better scores in the *TagMF* condition ($M = 4.22$, $SD = 0.89$) than in the tag-based one ($M = 4.15$, $SD = 0.82$). Admittedly, the effect size was small ($d = 0.09$), and with a two-tailed t -test we found no evidence for a significant difference ($t(52) = 0.32$, $p = .752$). This, however, confirms H6.

Critiquing. Regarding the critiquing process, and in particular, the tags we used as critique dimensions in our prototype, a MANOVA aggregating several questionnaire items taken from [18] indicated no significant difference between conditions, $F(12, 41) = 0.68$, $p = .761$, $\eta_p^2 = .17$. Table 4 shows the individual results for these items, which were assessed once, after participants completed all tasks.

Overall, we found that participants understood the critique dimensions and their effect on the results. Moreover, they liked to apply critiques in form of user-generated tags to influence the recommendation process. Considering qualitative feedback, one participant, for instance, answered to the open-ended question that it was “clear and straight-forward to point the system in the direction of movies he or she would like to watch”. However, others commented that it “would have been helpful to see a list of all tags as it was difficult to come up with suitable ones” (note that autocompletion was provided) and that

Table 4: *t*-test results with mean values and standard deviations ($df = 52$) comparing the conditions with respect to the tags used as critique dimensions (d represents Cohen’s effect size value). Higher values indicate better results.

Questionnaire Item	Tag-based		TagMF		<i>T</i>	<i>p</i>	<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
The tags made sense to me	4.22	0.75	4.48	0.75	1.27	.106	0.35
The tags helped me learn about the movie	4.00	0.73	4.26	0.76	1.27	.105	0.35
I like having the ability to specify critiques	4.52	0.64	4.67	0.68	0.82	.207	0.23
Movies displayed in response to my critique made sense	3.67	1.04	3.89	1.12	0.76	.227	0.21

they “missed a broader range of tags to select from”. Still, the questionnaire results were very positive in both conditions, with mean values even being slightly higher for *TagMF*. In summary, in spite of the lack of significances (according to the one-tailed *t*-tests we conducted, see Table 4) and of high effect sizes, we can thus at least partly accept H7, especially considering the minor (and only algorithmic) differences between conditions with respect to the critique dimensions.

Usability. In line with our more specific hypotheses, we used a one-tailed *t*-test to analyze whether participants were in general more satisfied in the *TagMF* condition: Results indicated a higher satisfaction ($M = 4.48$, $SD = 0.75$) with the corresponding system variant than in the control group ($M = 4.11$, $SD = 0.80$), with significant difference and medium effect size ($t(52) = 1.75$, $p = .043$, $d = 0.48$). One participant in the *TagMF* condition, for example, explicitly stated the he or she “enjoyed using the system”.

More concretely, usability of the two variants of our prototype system was rated equally “good”, with a SUS score of 87 in the *TagMF* condition, and 84 in the other. A two-tailed *t*-test showed no significant difference ($t(52) = 1.12$, $p = .269$) and only a rather small effect ($d = .30$). This corresponds to the very positive assessment of interface adequacy in both the *TagMF* condition ($M = 4.44$, $SD = 0.57$) and the tag-based one ($M = 4.20$, $SD = 0.53$), without significant difference ($t(52) = 1.55$, $p = .128$) and medium effect size ($d = .44$). Regarding the UEQ, values between 1.34 and 2.43 on the different subscales were very promising for *TagMF*, as shown in Table 5. In particular, subscales for transparency and efficiency yielded “excellent” scores, and control was rated as “good”. Overall, scores were inferior in the tag-based condition, with values in a range from 1.18 to 2.20. Efficiency was only rated as “good” and control as “above average”. In terms of control and stimulation, two-tailed *t*-tests even indicated significant differences with medium effect size.

5.3.4. Discussion

In the second user study, we examined how *TagMF* can be exploited for integrating model-based CF with interac-

Table 5: *t*-test results with mean values and standard deviations ($df = 52$) comparing the conditions with respect to the UEQ subscales (d represents Cohen’s effect size value). Higher values indicate better results on the 7-point bipolar scale.

Subscale	Tag-based		TagMF		<i>T</i>	<i>p</i>	<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Attractiveness	1.73	0.65	1.99	0.79	1.28	.206	0.35
Transparency	2.20	0.63	2.43	0.61	1.32	.194	0.36
Efficiency	1.70	0.62	1.95	0.63	1.47	.148	0.40
Control	1.22	0.58	1.61	0.75	2.13	.038*	0.58
Stimulation	1.36	0.66	1.78	0.80	2.09	.042*	0.57
Novelty	1.18	0.94	1.34	1.09	0.60	.550	0.16

tive critiquing, taking the critiques applied to the currently recommended item and, in contrast to typical critique-based RS, the user’s existing long-term preference profile into account. First, we would like to draw attention on the very positive assessment of the movie chosen as a starting point after completing the preliminary task. Participants were asked to select this movie from the initial set of recommendations notably generated by means of *TagMF* in both conditions. The results corroborate findings from the first user study showing that our method indeed leads to very adequate suggestions (see Section 5.2.3).

After each of the main tasks, we obtained very promising results regarding perception of recommendation quality (H1). With exception of Task 2a, differences were significant¹⁹. When participants had to find movies fitting their personal interests, i.e. especially in Task 1, the value of *TagMF* for the critiquing process became even more apparent: The underlying CF model allows to consider preference profiles (i.e. user-factor vectors) learned over a potentially longer period of time via conventional preference elicitation. Thus, suggestions are more likely to correspond not only to critiques applied due to situational aspects of the search process, but to the user’s general interests as it is typical for CF recommenders. This positive assessment of the recommendations is reflected in the scores for the more specific constructs, mean item rating and choice satisfaction, which are higher for *TagMF* in all cases, most often significantly (H1, H2).

The positive impact latent knowledge has on the critiquing process and resulting recommendations compared to when only (user-generated) content information serves as input data (as it is customary in critique-based RS), is also supported by other relevant variables related to subjective system aspects and user experience. For instance, while purely content-based approaches are known to tend to over-specialization [81], i.e. recommending similar items, we found significant or at least marginal improvements regarding diversity of recommended item sets due to using our method (H5). This is well in line with other works that propose to exploit latent factors to di-

¹⁹Potentially because it was harder for participants to determine whether recommended items fitted the goal of the task than in the two other tasks (as interests of the fictitious date had to be taken into account), the difference here was only marginal.

verify RS results or to address the “filter bubble” problem [e.g. 58, 60]. Concerning choice difficulty, we in contrast did not find a positive effect: The prototype variant that reimplemented the method behind *MovieTuner* even tended to make it easier to choose a movie from the set of recommendations²⁰. As a consequence, while we assumed that taking long-term interests into account by means of *TagMF* would make it more easy to decide, we have to reject H3. However, usability assessment of the different system variants indicated no negative impact on user experience in general. As in the first user study, most usability-related comments were independent of the respective condition: In their qualitative feedback, participants wanted, for instance, to “directly search for movies” or to “exclude bad movies and keep good movies over several critiquing cycles”. Consequently, we will address these more general aspects in future iterations of our prototype system, although they are actually more related to system use in real-world scenarios.

With respect to transparency, we found only marginal improvements due to using *TagMF*. Bearing in mind that in this case latent information comes into play, the results however even shed a positive light (H6): It would not have been surprising if the variant that exclusively relied on well understandable tag-based information had facilitated the comprehension of recommendations. In principle, the same applies to perceived effort and the more objective measurement, time spent for tasks. Yet, we assumed that considering the user’s preference profile would have a positive effect on his or her efficiency when navigating through the information space. As the results were however only slightly better with our method, we have to reject H4.

Besides aspects related to recommendations, we investigated the effect of *TagMF* on the selection of the critique dimensions that served as a means to take participants’ interests as well as task-related goals into account. Overall, participants expressed more positive feedback. Differences between conditions were not significant, but we blended together tags chosen according to our method with the ones selected based on item-tag relevance data to equal proportions, i.e. only 3 of the user-generated tags shown as critique dimensions were actually determined differently. This minor difference in the user interface, in combination with the between-subject design, might have diminished the effect of taking the CF profile into account. Effect sizes were still small to nearly medium, but it has to be noted that explanatory power was limited due to the number of participants. However, participants were confronted with the related questionnaire items only once, after completing all tasks. Thus, tasks where participants in addition to their personal preferences had to consider interests of

others might have distorted the results: In these cases, personalized critique dimensions specifically tailored towards individual long-term interests by means of *TagMF* might indeed been less useful. The answers to the open-ended question support this assumption. For instance, one participant mentioned that it was “difficult to quickly change the direction of recommendations (from horror to comedy) in order to obtain movies for a 9-year-old”. Yet, he or she explicitly added that “adapting to the user profile is, on the other hand, purpose of the system”. Unfortunately, in contrast to the first user study (see Section 5.2.4), SEM did not lead to meaningful insights because of sample size and study design. As a consequence, we plan to further investigate how employing our method may affect the subjective assessment of critiquing, and thus user experience, with a larger number of participants. Overall, we can still at least partly accept H7.

In general, participants in the *TagMF* condition stated to be more satisfied than in the tag-based condition, with significant difference. Taken all together, enhancing model-based CF according to our method can thus be seen as a promising means to add interaction possibilities and to improve user experience. This validates that *TagMF* can be successfully applied as an extension as described in Section 4.3 to allow users critiquing a recommended item in this typically very restricted type of RS (RQ1c).

For specifically examining the value of learning a latent factor model that additionally integrates user-generated tags, the second user study was designed as a comparison with an interactive RS that similar to the well-known *MovieTuner* relied on an entirely tag-based model. As already outlined above, results were overall very positive: We observed that using *TagMF* led to significantly better recommendations in terms of subjective aspects such as perceived quality and diversity. The positive results are reflected in user experience, e.g. choice satisfaction, and in the higher average ratings provided for the items eventually recommended after finishing the critiquing process. On the other hand, particularly our usability evaluation yielded only slightly better results. Given sample size, the small visible differences between the prototype variants, and the potential confusion that might be induced by the latent factors in the *TagMF* condition, the absence of differences, however, already appears promising. Nonetheless, further investigation and larger user studies are required in this area. In summary, from a user perspective, considering additional content data according to our method yet seems beneficial in comparison to the wide range of interactive recommending approaches that solely rely on (user-generated) content information (see Section 2.1) and, as a consequence, cannot consider user profiles based on past user-item interaction data (RQ2b).

Finally, observing participants’ behavior indicated that they valued that in the prototype variant based on *TagMF*, a tag cloud allowed to inspect their formerly opaque representation within the underlying model. The successful implementation of such a tag-based explanation in our pro-

²⁰Note that in the first user study, we additionally assessed this variable objectively by measuring how long it took participants to settle on a recommended movie. Due to differences in study setup and task descriptions, it was only possible to assess this construct in a subjective manner for the second user study.

prototype system according to Section 4.4 shows how additional information may be used in typical model-based CF systems for explaining existing long-term preference profiles (RQ1d). However, both user studies have not been focused on this aspect, making it subject of future work to more completely validate this application possibility. Concretely, we plan to conduct another empirical study to investigate actual comprehensibility of the tag cloud and to compare ours with other approaches that explain recommendations, in particular ones that use tags, e.g. [82].

6. Conclusions and Outlook

In this paper, we have introduced *TagMF*, a method that combines the benefits of latent factors derived by standard MF with the ones of user-generated tags. As discussed in Section 2, MF is an efficient means for generating precise recommendations and has been improved by several algorithmic advances in the last years, for instance, by enhancing the factor models with additional information. However, accuracy improvements as measured in retrospective offline experiments have not always contributed to user satisfaction to the same extent. Interactive recommending approaches, which have been shown to increase the level of user control and system transparency, in contrast, often use entirely different algorithms, thus being independent of the advantages provided e.g. by state-of-the-art model-based CF techniques.

Following our research questions posed at the beginning of this paper in Section 1, we have shown the value of additional content information such as tags when used in CF: Integrating item-related tag relevance data by means of *TagMF* allows to derive corresponding user-related tag relevance data (i.e. we do not require up front availability of tags describing the current user’s interest, but instead infer this information) as well as tag-factor relations. This contributes to increased recommendation quality and simultaneously opens novel ways to extend typical automated RS with interactive techniques, thereby overcoming several of their widely discussed drawbacks. Users can be offered more control over the recommendation process, which is in contemporary real-world systems usually limited to (re-)rating single items. Concretely, the application possibilities of *TagMF* allow users to interactively adapt the set of items suggested as known from standard MF towards their current needs and goals through easily comprehensible tags—in cold-start situations (RQ1a), with an existing profile (RQ1b), and by critiquing a recommended item (RQ1c)—and to inspect their long-term preference profile with the aid of tag-based explanations (RQ1d).

The offline experiments we conducted corroborate that *TagMF* increases objective recommendation quality (see Section 5.1.2). Yet, it has to be noted that additional parameter tuning is necessary, i.e. the number of tags to be taken into account must be determined, and that model learning becomes more complex. On the other hand, a qualitative inspection performed in this context underlines

that the method is able to reveal inherent meanings of the resulting, usually abstract latent factor models by incorporating the easy-to-understand semantics of user-generated tags (see Section 5.1.3). Still, one must also note that this data needs to be collected—or other datasources must be available—before being able to apply our method. Two quantitative user studies with an interactive web-based prototype movie RS served to validate the application possibilities proposed in Section 4, which are directly related to our research questions. While participants were not allowed to create tags themselves, we believe this involves no loss of generality as the well-known dataset we used consists of a very large set of tags generated by the user community of a similar system. Besides, in a real-world scenario, it would be possible to easily apply *TagMF* with any kind of additional data. Consequently, these studies together can be considered to constitute the first extensive empirical evaluation in RS research with respect to the use of additional information in CF.

The first user study presented in Section 5.2 confirmed for the first time a positive influence on the subjective assessment of system aspects, as well as on user experience in general. In particular, perceived recommendation quality and transparency benefited from the integrated *TagMF* model. As a consequence, participants were able to decide faster and were more satisfied with their chosen item. Interestingly, besides the fact that they liked the interaction via tags generally more, results were especially promising with respect to the elicitation of initial preferences. Apparently, integrating our method seems to be quite useful in cold-start situations, as selecting a small number of tags led to recommendations at least as good as rating a larger number of items ex ante. Using SEM, we further analyzed these findings, focusing on the role of transparency and the impact of different preference elicitation methods. In this way, overall, the first study allowed us to validate the application possibilities described in Section 4.1 and 4.2, referring to RQ1a and RQ1b, and being focused on a comparison with an automated rating-based MF system, to answer RQ2a.

The second user study presented in Section 5.3 shows the value of considering user-generated content in addition to latent knowledge in interactive recommending scenarios: The results emphasize that using *TagMF*, a personalized critique-based recommendation process can successfully be integrated as an extension to standard model-based CF systems. We again obtained positive feedback with respect to subjective system aspects, e.g. perceived recommendation quality and diversity, and regarding constructs related to user experience, e.g. choice satisfaction. Note that while the number of participants was limited in consideration of the study’s between-subject design, the effect sizes in addition to statistical significances generally confirm the benefits of our method. Nevertheless, experiments with more participants would be required to reach significance more often and to make even stronger claims. Yet, especially given the minor differences between the condi-

tions, we believe that the current results already provide sufficient evidence in favor of our method. Overall, the second study therefore allowed us to validate the application possibility described in Section 4.3, referring to RQ1c, and to complement the user-centric evaluation against state-of-the-art recommending approaches by a comparison with a purely tag-based interactive recommender, thus answering RQ2b.

In summary, enhancing a recommender according to our method appears to be a promising means to provide additional interactive features in today’s automated systems and to increase their transparency. We successfully addressed RQ1 by describing several application possibilities of *TagMF*, which we validated in our user studies. Qualitative comments of participants (see Section 5.2 and 5.3) suggest that there are indeed usability-related aspects of our prototype system that could be improved. Consequently, although more related to real-world use, these issues are of interest for future work and will then be further investigated by means of, among others, qualitative methods. In general, we still received very positive results with respect to usability, and in particular the novel interactive features that can be integrated in CF systems by using *TagMF*. In this context, it has to be noted that although it could have been expected, the extended interaction mechanisms had no negative influence on e.g. perceived effort. With regard to RQ2, the user studies allowed us to investigate the effect applying our method has on subjective system aspects and user experience in comparison to established baselines: Learning an integrated model of latent factors and additional (user-generated) information such as tags led to significantly better scores in a majority of comparisons, emphasizing the value of *TagMF* for implementing interactive RS.

In future work, we plan to exploit the integration of user-generated tags, other content-related information (e.g. metadata on genres or keywords extracted from social media) and contextual attributes that likely affect the consumption experience (such as the current season when recommending e.g. Christmas movies) into MF more extensively. The effects of using different kinds of data as well as of enriching other recommendation methods such as deep learning with additional data still need to be investigated. In doing so, we also aim at improving current as well as developing novel application possibilities for *TagMF*. For instance, one can think of more advanced interaction mechanisms as well as improved (and possibly visually-enhanced) explanations. In line with this, we are interested in conducting further empirical user studies focusing, among others, on our tag-based explanations: As of now, we have answered RQ1d by describing in Section 4.4 a way additional content information can be used in model-based CF for explaining an existing preference profile. The derived user-tag relations should by construction describe the latent part of the user profile in an adequate manner. This is supported by the qualitative inspection we performed on the latent factor space, the successful

implementation of the tag cloud in our prototype system, and the observation of participants’ behavior. However, as the two present studies had a different focus, a more profound validation of this application possibility is left for future work. Moreover, although *TagMF* can be considered easily usable with other items than movies due to the domain independence of the underlying CF and the small additional requirements of our method, we especially want to evaluate it when applied to other domains.

References

- [1] X. Su, T. M. Khoshgoftaar, A survey of collaborative filtering techniques, *Advances in Artificial Intelligence 2009 (2009)* 4:1–4:19. doi:10.1155/2009/421425.
- [2] F. Ricci, L. Rokach, B. Shapira, *Recommender Systems Handbook*, Springer US, Boston, MA, USA, 2015, Ch. Recommender Systems: Introduction and Challenges, pp. 1–34. doi:10.1007/978-1-4899-7637-6_1.
- [3] G. Jawaheer, P. Weller, P. Kostkova, Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback, *ACM Transactions on Interactive Intelligent Systems* 4 (2) (2014) 8:1–8:26. doi:10.1145/2512208.
- [4] A. Gunawardana, G. Shani, *Recommender Systems Handbook*, Springer US, Boston, MA, USA, 2015, Ch. Evaluating Recommender Systems, pp. 265–308. doi:10.1007/978-1-4899-7637-6_8.
- [5] B. Xiao, I. Benbasat, E-commerce product recommendation agents: Use, characteristics, and impact, *MIS Quarterly* 31 (1) (2007) 137–209.
- [6] J. A. Konstan, J. Riedl, Recommender systems: From algorithms to user experience, *User Modeling and User-Adapted Interaction* 22 (1-2) (2012) 101–123. doi:10.1007/s11257-011-9112-x.
- [7] P. Pu, L. Chen, R. Hu, Evaluating recommender systems from the user’s perspective: Survey of the state of the art, *User Modeling and User-Adapted Interaction* 22 (4-5) (2012) 317–355. doi:10.1007/s11257-011-9115-7.
- [8] M. Ge, C. Delgado-Battenfeld, D. Jannach, Beyond accuracy: Evaluating recommender systems by coverage and serendipity, in: *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys ’10)*, ACM, New York, NY, USA, 2010, pp. 257–260. doi:10.1145/1864708.1864761.
- [9] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys ’11)*, ACM, New York, NY, USA, 2011, pp. 109–116. doi:10.1145/2043932.2043955.
- [10] M. Jugovac, D. Jannach, Interacting with recommenders – Overview and research directions, *ACM Transactions on Interactive Intelligent Systems* 7 (3) (2017) 10:1–10:46. doi:10.1145/3001837.
- [11] G. Linden, B. Smith, J. York, Amazon.com recommendations: Item-to-item collaborative filtering, *IEEE Internet Computing* 7 (1) (2003) 76–80. doi:10.1109/MIC.2003.1167344.
- [12] J. Bennett, S. Lanning, The Netflix prize, in: *Proceedings of the KDD Cup and Workshop*, ACM, New York, NY, USA, 2007, pp. 3–6.
- [13] E. Pariser, *The Filter Bubble: What the Internet is Hiding From You*, Penguin Press, 2011.
- [14] S. Nagulendra, J. Vassileva, Understanding and controlling the filter bubble through interactive visualization: A user study, in: *Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT ’14)*, ACM, 2014, pp. 107–115. doi:10.1145/2631775.2631811.
- [15] N. Tintarev, J. Masthoff, *Recommender Systems Handbook*, Springer US, Boston, MA, USA, 2015, Ch. Explaining Recommendations: Design and Evaluation, pp. 353–382. doi:10.1007/978-1-4899-7637-6_10.

- [16] B. P. Knijnenburg, M. C. Willemsen, *Recommender Systems Handbook*, Springer US, Boston, MA, USA, 2015, Ch. Evaluating Recommender Systems with User Experiments, pp. 309–352. doi:10.1007/978-1-4899-7637-6_9.
- [17] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, C. Newell, Explaining the user experience of recommender systems, *User Modeling and User-Adapted Interaction* 22 (4-5) (2012) 441–504. doi:10.1007/s11257-011-9118-4.
- [18] J. Vig, S. Sen, J. Riedl, Navigating the tag genome, in: *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI '11)*, ACM, New York, NY, USA, 2011, pp. 93–102. doi:10.1145/1943403.1943418.
- [19] S. Sen, J. Vig, J. Riedl, Tagommenders: Connecting users to items through tags, in: *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*, ACM, New York, NY, USA, 2009, pp. 671–680. doi:10.1145/1526709.1526800.
- [20] Z. Guan, C. Wang, J. Bu, C. Chen, K. Yang, D. Cai, X. He, Document recommendation in social tagging services, in: *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*, ACM, New York, NY, USA, 2010, pp. 391–400. doi:10.1145/1772690.1772731.
- [21] L. Chen, P. Pu, Critiquing-based recommenders: Survey and emerging trends, *User Modeling and User-Adapted Interaction* 22 (1-2) (2012) 125–150. doi:10.1007/s11257-011-9108-6.
- [22] S. Bostandjiev, J. O'Donovan, T. Höllerer, TasteWeights: A visual interactive hybrid recommender system, in: *Proceedings of the 6th ACM Conference on Recommender Systems (RecSys '12)*, ACM, New York, NY, USA, 2012, pp. 35–42. doi:10.1145/2365952.2365964.
- [23] D. Parra, P. Brusilovsky, C. Trattner, See what you want to see: Visual user-driven approach for hybrid recommendation, in: *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI '14)*, ACM, New York, NY, USA, 2014, pp. 235–240. doi:10.1145/2557500.2557542.
- [24] Y. Koren, R. M. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, *IEEE Computer* 42 (8) (2009) 30–37. doi:10.1109/MC.2009.263.
- [25] Y. Koren, R. M. Bell, *Recommender Systems Handbook*, Springer US, Boston, MA, USA, 2015, Ch. Advances in Collaborative Filtering, pp. 77–118. doi:10.1007/978-1-4899-7637-6_3.
- [26] T. Donkers, B. Loepp, J. Ziegler, Towards understanding latent factors and user profiles by enhancing matrix factorization with tags, in: *Poster Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*, 2016.
- [27] A. Karatzoglou, X. Amatriain, L. Baltrunas, N. Oliver, Multi-verse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering, in: *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys '10)*, ACM, New York, NY, USA, 2010, pp. 79–86. doi:10.1145/1864708.1864727.
- [28] P. Forbes, M. Zhu, Content-boosted matrix factorization for recommender systems: Experiments with recipe recommendation, in: *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11)*, ACM, New York, NY, USA, 2011, pp. 261–264. doi:10.1145/2043932.2043979.
- [29] J. McAuley, J. Leskovec, Hidden factors and hidden topics: Understanding rating dimensions with review text, in: *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*, ACM, New York, NY, USA, 2013, pp. 165–172. doi:10.1145/2507157.2507163.
- [30] Y. Shi, M. Larson, A. Hanjalic, Mining contextual movie similarity with matrix factorization for context-aware recommendation, *ACM Transactions on Intelligent Systems and Technology* 4 (1) (2013) 16:1–16:19. doi:10.1145/2414425.2414441.
- [31] Q. Diao, M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, C. Wang, Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS), in: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*, ACM, New York, NY, USA, 2014, pp. 193–202. doi:10.1145/2623330.2623758.
- [32] I. Fernández-Tobías, I. Cantador, Exploiting social tags in matrix factorization models for cross-domain collaborative filtering, in: *Proceedings of the 1st Workshop on New Trends in Content-based Recommender Systems (CBRecSys '14)*, 2014, pp. 34–41.
- [33] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, S. Ma, Explicit factor models for explainable recommendation based on phrase-level sentiment analysis, in: *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14)*, ACM, New York, NY, USA, 2014, pp. 83–92. doi:10.1145/2600428.2609579.
- [34] A. Almahairi, K. Kastner, K. Cho, A. Courville, Learning distributed representations from reviews for collaborative filtering, in: *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15)*, ACM, New York, NY, USA, 2015, pp. 147–154. doi:10.1145/2792838.2800192.
- [35] J. Nguyen, M. Zhu, Content-boosted matrix factorization techniques for recommender systems, *Statistical Analysis and Data Mining* 6 (4) (2013) 286–301. doi:10.1002/sam.11184.
- [36] B. Muthén, A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika* 49 (1) (1984) 115–132. doi:10.1007/BF02294210.
- [37] E. I. Sparling, S. Sen, Rating: How difficult is it?, in: *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11)*, ACM, New York, NY, USA, 2011, pp. 149–156. doi:10.1145/2043932.2043961.
- [38] P. Cremonesi, F. Garzotto, R. Turrin, User effort vs. accuracy in rating-based elicitation, in: *Proceedings of the 6th ACM Conference on Recommender Systems (RecSys '12)*, ACM, New York, NY, USA, 2012, pp. 27–34. doi:10.1145/2365952.2365963.
- [39] D. Parra, X. Amatriain, Walk the talk: Analyzing the relation between implicit and explicit feedback for preference elicitation, in: *Proceedings of the 19th International Conference on User Modeling, Adaptation and Personalization (UMAP '11)*, Springer, Berlin, Germany, 2011, pp. 255–268. doi:10.1007/978-3-642-22362-4_22.
- [40] C. He, D. Parra, K. Verbert, Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities, *Expert Systems with Applications* 56 (1) (2016) 9–27. doi:10.1016/j.eswa.2016.02.013.
- [41] B. Loepp, C.-M. Barbu, J. Ziegler, Interactive recommending: Framework, state of research and future challenges, in: *Proceedings of the 1st Workshop on Engineering Computer-Human Interaction in Recommender Systems (EnCHIREs '16)*, 2016, pp. 3–13.
- [42] M. Mandl, A. Felfernig, Improving the performance of unit critiquing, in: *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP '12)*, Springer, Berlin, Germany, 2012, pp. 176–187. doi:10.1007/978-3-642-31454-4_15.
- [43] B. Loepp, K. Herrmann, J. Ziegler, Blended recommending: Integrating interactive information filtering and algorithmic recommender techniques, in: *Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems (CHI '15)*, ACM, New York, NY, USA, 2015, pp. 975–984. doi:10.1145/2702123.2702496.
- [44] C. di Sciascio, V. Sabol, E. E. Veas, Rank as you go: User-driven exploration of search results, in: *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*, ACM, New York, NY, USA, 2016, pp. 118–129. doi:10.1145/2856767.2856797.
- [45] I. Andjelkovic, D. Parra, J. O'Donovan, Moodplay: Interactive mood-based music discovery and recommendation, in: *Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '16)*, ACM, New York, NY, USA, 2016, pp. 275–279. doi:10.1145/2930238.2930280.
- [46] K. Verbert, D. Parra, P. Brusilovsky, Agents vs. users: Visual recommendation of research talks with multiple dimension of relevance, *ACM Transaction on Interactive Intelligent Systems* 6 (2) (2016) 11:1–11:42. doi:10.1145/2946794.

- [47] T. T. Nguyen, J. Riedl, Predicting users' preference from tag relevance, in: Proceedings of the 21st International Conference on User Modeling, Adaptation and Personalization (UMAP '13), Springer, Berlin, Germany, 2013, pp. 274–280. doi:10.1007/978-3-642-38844-6_23.
- [48] G. E. Forsythe, M. A. Malcolm, C. B. Moler, Computer Methods for Mathematical Computations, Prentice Hall, Englewood Cliffs, NJ, USA, 1977, Ch. Least Squares and the Singular Value Decomposition.
- [49] S. Rendle, L. Schmidt-Thieme, Online-updating regularized kernel matrix factorization models for large-scale recommender systems, in: Proceedings of the 2nd ACM Conference on Recommender Systems (RecSys '08), ACM, New York, NY, USA, 2008, pp. 251–258. doi:10.1145/1454008.1454047.
- [50] H. Wang, N. Wang, D.-Y. Yeung, Collaborative deep learning for recommender systems, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15), ACM, New York, NY, USA, 2015, pp. 1235–1244. doi:10.1145/2783258.2783273.
- [51] T. V. Nguyen, A. Karatzoglou, L. Baltrunas, Gaussian process factorization machines for context-aware recommendations, in: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14), ACM, New York, NY, USA, 2014, pp. 63–72. doi:10.1145/2600428.2609623.
- [52] K. Zhou, S.-H. Yang, H. Zha, Functional matrix factorizations for cold-start recommendation, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11), ACM, New York, NY, USA, 2011, pp. 315–324. doi:10.1145/2009916.2009961.
- [53] R. Karimi, C. Freudenthaler, A. Nanopoulos, L. Schmidt-Thieme, Exploiting the characteristics of matrix factorization for active learning in recommender systems, in: Proceedings of the 6th ACM Conference on Recommender Systems (RecSys '12), ACM, New York, NY, USA, 2012, pp. 317–320. doi:10.1145/2365952.2366031.
- [54] X. Zhao, W. Zhang, J. Wang, Interactive collaborative filtering, in: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM '13), ACM, New York, NY, USA, 2013, pp. 1411–1420. doi:10.1145/2505515.2505690.
- [55] M. Elahi, F. Ricci, N. Rubens, A survey of active learning in collaborative filtering recommender systems, Computer Science Review 20 (2016) 29–50. doi:10.1016/j.cosrev.2016.05.002.
- [56] B. Loepp, T. Hussein, J. Ziegler, Choice-based preference elicitation for collaborative filtering recommender systems, in: Proceedings of the 32nd ACM Conference on Human Factors in Computing Systems (CHI '14), ACM, New York, NY, USA, 2014, pp. 3085–3094. doi:10.1145/2556288.2557069.
- [57] M. P. Graus, M. C. Willemsen, Improving the user experience during cold start through choice-based preference elicitation, in: Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15), ACM, New York, NY, USA, 2015, pp. 273–276. doi:10.1145/2792838.2799681.
- [58] M. C. Willemsen, M. P. Graus, B. P. Knijnenburg, Understanding the role of latent feature diversification on choice difficulty and satisfaction, User Modeling and User-Adapted Interaction 26 (4) (2016) 347–389. doi:10.1007/s11257-016-9178-6.
- [59] E. Gansner, Y. Hu, S. Kobourov, C. Volinsky, Putting recommendations on the map: Visualizing clusters and relations, in: Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys '09), ACM, New York, NY, USA, 2009, pp. 345–348. doi:10.1145/1639714.1639784.
- [60] J. Kunkel, B. Loepp, J. Ziegler, A 3D item space visualization for presenting and manipulating user preferences in collaborative filtering, in: Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17), ACM, New York, NY, USA, 2017, pp. 3–15. doi:10.1145/3025171.3025189.
- [61] B. Németh, G. Takács, I. Pilászy, D. Tikk, Visualization of movie features in collaborative filtering, in: Proceedings of the 12th IEEE International Conference on Intelligent Software Methodologies, Tools and Techniques (SoMeT '13), IEEE, Washington, DC, USA, 2013, pp. 229–233. doi:10.1109/SoMeT.2013.6645674.
- [62] M. Rossetti, F. Stella, M. Zanker, Towards explaining latent factors with topic models in collaborative recommender systems, in: Proceedings of the 24th International Workshop on Database and Expert Systems Applications (DEXA '13), 2013, pp. 162–167. doi:10.1109/DEXA.2013.26.
- [63] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, J. A. Konstan, User perception of differences in recommender algorithms, in: Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14), ACM, New York, NY, USA, 2014, pp. 161–168. doi:10.1145/2645710.2645737.
- [64] D. Bollen, B. P. Knijnenburg, M. C. Willemsen, M. P. Graus, Understanding choice overload in recommender systems, in: Proceedings of the 4th ACM Conference on Recommender Systems (RecSys '10), ACM, New York, NY, USA, 2010, pp. 63–70. doi:10.1145/1864708.1864724.
- [65] T. Donkers, B. Loepp, J. Ziegler, Merging latent factors and tags to increase interactive control of recommendations, in: Poster Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15), 2015.
- [66] E. H. Moore, On the reciprocal of the general algebraic matrix, Bulletin of the American Mathematical Society 26 (1920) 394–395. doi:10.1090/S0002-9904-1920-03322-7.
- [67] R. Penrose, A generalized inverse for matrices, Mathematical Proceedings of the Cambridge Philosophical Society 51 (3) (1955) 406–413. doi:10.1017/S0305004100030401.
- [68] G. Takács, I. Pilászy, B. Németh, D. Tikk, Scalable collaborative filtering approaches for large recommender systems, Journal of Machine Learning Research 10 (2009) 623–656.
- [69] A. Said, A. Bellogín, RiVal: A toolkit to foster reproducibility in recommender system evaluation, in: Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14), ACM, New York, NY, USA, 2014, pp. 371–372. doi:10.1145/2645710.2645712.
- [70] T. Donkers, B. Loepp, J. Ziegler, Tag-enhanced collaborative filtering for increasing transparency and interactive control, in: Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '16), ACM, New York, NY, USA, 2016, pp. 169–173. doi:10.1145/2930238.2930287.
- [71] J. Vig, S. Sen, J. Riedl, Computing the tag genome, Tech. rep., University of Minnesota (2010).
- [72] B. P. Knijnenburg, M. C. Willemsen, A. Kobsa, A pragmatic procedure to support the user-centric evaluation of recommender systems, in: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11), ACM, New York, NY, USA, 2011, pp. 321–324. doi:10.1145/2043932.2043993.
- [73] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11), ACM, New York, NY, USA, 2011, pp. 157–164. doi:10.1145/2043932.2043962.
- [74] J. Brooke, SUS – A quick and dirty usability scale, in: Usability Evaluation in Industry, Taylor & Francis, London, UK, 1996, pp. 189–194.
- [75] B. Laugwitz, T. Held, M. Schrepp, Construction and evaluation of a user experience questionnaire, in: Proceedings of the 4th Symposium of the Austrian HCI and Usability Engineering Group (USAB '08), Springer, Berlin, Germany, 2008, pp. 63–76. doi:10.1007/978-3-540-89350-9_6.
- [76] B. P. Knijnenburg, M. C. Willemsen, The effect of preference elicitation methods on the user experience of a recommender system, in: Extended Abstracts of the 28th ACM Conference on Human Factors in Computing Systems (CHI '10), ACM, New York, NY, USA, 2010, pp. 3457–3462. doi:10.1145/1753846.1754001.
- [77] T. T. Nguyen, D. Kluver, T.-Y. Wang, P.-M. Hui, M. D. Ekstrand, M. C. Willemsen, J. Riedl, Rating support interfaces to improve user experience and recommender accuracy, in: Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13), ACM, New York, NY, USA, 2013, pp. 149–156. doi:10.1145/2507157.2507188.

- [78] M. D. Ekstrand, D. Kluver, F. M. Harper, J. A. Konstan, Letting users choose recommender algorithms: An experimental study, in: Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15), ACM, New York, NY, USA, 2015, pp. 11–18. doi:10.1145/2792838.2800195.
- [79] J. Vig, S. Sen, J. Riedl, The tag genome: Encoding community knowledge to support novel interaction, ACM Transactions on Interactive Intelligent Systems 2 (3) (2012) 13:1–13:44. doi:10.1145/2362394.2362395.
- [80] A. M. Rashid, I. Albert, D. Cosley, S. K. Lam, S. M. McNee, J. A. Konstan, J. Riedl, Getting to know you: Learning new user preferences in recommender systems, in: Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI '02), ACM, New York, NY, USA, 2002, pp. 127–134. doi:10.1145/502716.502737.
- [81] L. Iaquina, M. d. Gemmis, P. Lops, G. Semeraro, M. Filannino, P. Molino, Introducing serendipity in a content-based recommender system, in: Proceedings of the 8th International Conference on Hybrid Intelligent Systems (HIS '08), IEEE, Washington, DC, USA, 2008, pp. 168–173. doi:10.1109/HIS.2008.25.
- [82] J. Vig, S. Sen, J. Riedl, Tagsplanations: Explaining recommendations using tags, in: Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09), ACM, New York, NY, USA, 2009, pp. 47–56. doi:10.1145/1502650.1502661.

The following article is reused from:

Donkers, T., Kleemann, T., & Ziegler, J. (2020). Explaining recommendations by means of aspect-based transparent memories. In F. Paternò, & N. Oliver (Eds.), *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 166-176). The Association for Computing Machinery. <https://doi.org/10.1145/3377325.3377520>



Explaining Recommendations by Means of Aspect-based Transparent Memories

Tim Donkers, Timm Kleemann, Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
{firstname.lastname}@uni-due.de

ABSTRACT

Recommender Systems have seen substantial progress in terms of algorithmic sophistication recently. Yet, the systems mostly act as black boxes and are limited in their capacity to explain why an item is recommended. In many cases recommendations methods are employed in scenarios where users not only rate items, but also convey their opinion on various relevant aspects, for instance by the means of textual reviews. Such user-generated content can serve as a useful source for deriving explanatory information to increase system intelligibility and, thereby, the user's understanding.

We propose a recommendation and explanation method that exploits the comprehensiveness of textual data to make the underlying criteria and mechanisms that lead to a recommendation more transparent. Concretely, the method incorporates neural memories that store aspect-related opinions extracted from raw review data. We apply attention mechanisms to transparently write and read information from memory slots.

Besides customary offline experiments, we conducted an extensive user study. The results indicate that our approach achieves a higher overall quality of explanations compared to a state-of-the-art baseline. Utilizing Structural Equation Modeling, we additionally reveal three linked key factors that constitute explanation quality: Content adequacy, presentation adequacy, and linguistic adequacy.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **Human computer interaction (HCI)**;

KEYWORDS

Deep Learning; Recommender Systems; Explainable AI

ACM Reference Format:

Tim Donkers, Timm Kleemann, Jürgen Ziegler. 2020. Explaining Recommendations by Means of Aspect-based Transparent Memories. In *Twenty-fifth International Conference on Intelligent User Interfaces (IUI '20)*, March 17–20, 2020, Cagliari, Italy. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3377325.3377520>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

IUI '20, March 17–20, 2020, Cagliari, Italy
© 2020 Copyright is held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-7118-6/20/03...\$15.00
<https://doi.org/10.1145/3377325.3377520>

1 INTRODUCTION

Recommender Systems (RS) have grown quite mature in accurately suggesting items matching people's interest profiles. Although algorithmically sophisticated, the underlying methods oftentimes appear as black boxes incapable of providing any insights beyond Top- k rankings or predicted ratings. However, user satisfaction does not necessarily go in hand with system accuracy [22]. When lacking sufficient explanations, people may fail to develop trust and may ultimately reject the system's recommendations [15, 42].

When recommendation methods are applied in the real world, opaqueness can often be reduced by weaving in additional sources of information. Especially the consideration of human-generated content complements recommendations by providing insights conventional RS fail to disclose. For instance, textual reviews are known to have a strong influence on decision-making due to their comprehensiveness and information richness [2]. It is unsurprising that researchers suggest to design recommendation models conveying the rationale behind their decisions as well. Doing so can substantially increase transparency, comprehension, and trust [24, 33].

Review texts are a valuable source for a potentially transparent RS as they already contain statements that can be exploited to explain recommendations. Authors describe item properties and offer supporting or opposing arguments for certain aspects. From these propositions, the system can learn how to characterize users (and items) with respect to different dimensions. However, it is a challenging task to extract knowledge from such unstructured data because this would require, for instance, a general conception of semantics and structure of language.

Even if a system successfully exploits review texts for improving recommendations, this does not necessarily lead to more intelligible results, especially if the gained knowledge is still processed latently. Hence, an explainable RS approach should, in addition, maintain some of the review's original expressiveness and structure. This may not only help to generate explanations for recommendations but also to reflect the internal RS logic back to the user.

We propose a novel method called *Aspect-based Transparent Memories* (ATM) to model user preferences with respect to relevant aspects and compare them to an item's properties to predict ratings (see Figure 1). Neural memory-based methods [cf. 11, 40, 49] allow the externalization and structurization of possibly large amounts of knowledge. In our case, the memories are unique to a single person or item and control the process of encoding and decoding review data (see Section 3). Both steps, encoding (or *writing*) and decoding (or *reading*) are accompanied by mechanisms that are designed to impose transparency on the model. In the following, we describe the intuition behind our method in more detail, beginning with the memory components.

Reviews often contain very detailed information about what a user likes or dislikes. Therefore, reviews constitute a solid foundation for the representation of explicit knowledge about a certain user or item with respect to a set of aspects. However, not all parts of a review are equally important for all aspects. Hence, we distinguish between relevant and irrelevant sections by employing an attention mechanism [3, 26]. We act on the assumption that people do not randomly scatter information within a review but rather express thoughts and opinions in more or less fixed linguistic units. We have decided that the boundaries obtained by distinguishing sentences are a reasonable choice as they usually entail concrete statements about one or multiple aspects. Therefore, attention weights are distributed on sentence-level.

Although review data is highly informative, there still exist subtleties in a person's rating behavior that they might not even be aware of themselves. For instance, although never mentioned explicitly, a person might have a tendency for liking movies with strong female characters. It follows that information contained in reviews is still incomplete. As a consequence, in addition to the explicit representation gained from user reviews, we introduce an implicit representation that behaves similar to latent embeddings trained during conventional collaborative filtering [37]. Hence, ATM incorporates two distinct user states: their explicit interest in the aspects as well as the semantics still hidden in their ratings.

Now, in order to increase the explainability of the recommendation process, we match both states with knowledge about the target item. This knowledge is gained, again, by analyzing reviews, this time composed by other users about the target item. By doing so, ATM can identify concrete statements and judge whether they fit to a user's preferences, either explicit or implicit. Since people generally share a wide range of insights into a particular item, ATM can obtain very fine-grained explanations.

Put together, ATM is an explainable RS approach that exploits the power of neural memories to generate high quality recommendations in conjunction with explanations. We will empirically show that ATM outperforms a state-of-the-art review-based approach by running offline experiments on several datasets. Additionally, we will provide qualitative insights into the internal functioning of ATM. We, finally, present an extensive user study aimed at evaluating explanation quality as assessed by people.

The evaluation of textual explanations in RS research is often undertaken against overlap metrics such as BLEU [30]. However, offline evaluations miss the point of testing whether explanations actually help people make a decision. Only in some cases [e.g. 5, 45, 48], researchers have conducted user studies to assess the quality of explanations. But even in these exceptions, valuable insights that could be gained by considering psychological research concerning information processing and text comprehension are often disregarded.

Therefore, we have conducted a user study to identify which factors contribute to the generation of good textual explanations in a decision task and how these factors interact with each other. In a between-subjects design, we compared the performance of ATM explanations against a state-of-the-art review-retrieval method [5]. Statistical analyses based on Structural Equation Modeling reveal three central factors of textual explanation quality: content adequacy, presentation adequacy, and linguistic adequacy. By tracing

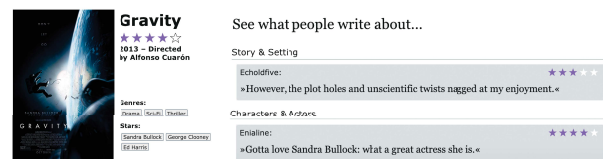


Figure 1: Screenshot of the ATM system with one extracted sentence per aspect.

back mediation paths through the structural model, we could verify that ATM achieves significantly better results concerning overall explanation quality. The greatest contributing factor to this finding is enhanced presentation adequacy. Hence, arranging textual explanations subject to salient aspects seems to be a more insightful way of presenting information than merely retrieving complete reviews. Summarized, our contributions are as follows:

- We present a novel memory-based recommendation approach designed to generate aspect-based explanations based on explicit and implicit knowledge about the user.
- We stress the importance of user studies in RS and Explainable AI research by presenting an extensive user study that provides insights into the generation of explanations.
- We analyze the study results based on powerful statistical tools to reveal three central interacting factors contributing to the quality of textual explanations: content adequacy, presentation adequacy, and linguistic adequacy.

2 RELATED WORK

Our work is related to several categories of research in RS and Artificial Intelligence. We draw motivation from findings in explainable RS and user-centered research. Methodically, ATM is connected to deep learning-based approaches, especially those utilizing user reviews to enhance recommendations. Specifically, there are connections to active areas of research that deal with developing model architectures based on attention and memory mechanisms. In the following, we give a short review of the mentioned research areas and distinguish our method from existing approaches.

One popular way of increasing transparency of RS is to use textual explanations [42]. When sufficient content information is available, item attributes may be aligned with user preferences to explain a recommendation [43]. Such data can also be used together with context information to point out arguments for recommended items, and may simultaneously serve as a means to critique recommendations [25]. For item-based collaborative filtering, the static variant used e.g. by Amazon (“Customers who bought this item also bought...”) is quite popular. For model-based methods, it is still difficult to improve transparency through explanations.

However, some approaches have emerged that try to achieve this. In [48], the authors exploit review data by identifying sentiments on a phrase-level to highlight product features the user is particularly interested in. Comparable works that built upon topic modeling can be found in [14, 34]. Specifically, all of the methods just mentioned follow a step-wise approach: First, they explicitly extract aspects and user opinions from phrases via sentiment analysis. Only afterwards, in the actual recommendation step, they generate

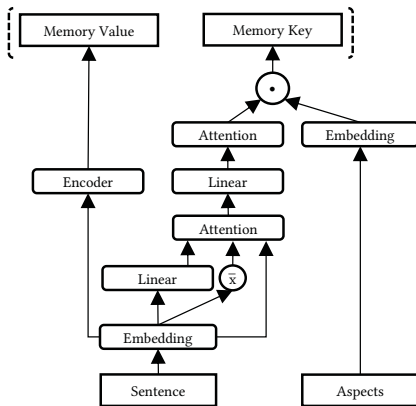


Figure 2: Schematic illustration of memory component calculation for a single sentence. \bar{x} indicates the mean of word embeddings in the sentence.

aspect-level explanations. As a result, the approaches are limited in that the aspect-extraction step is separated from the recommendation process. Additionally, as the authors in [50] point out, all of the above approaches define textual similarity as lexical similarity. What is ignored is the semantic similarity that can also occur with a low lexical overlap. Finally, the resulting explanations are extractions of words or phrases that become detached from their original context and may, therefore, be incomplete or distorted in meaning.

In contrast, we try to address these shortcoming by building an end-to-end neural network model with a strong focus in semantic similarity induced by embedding techniques [18, 28].

The authors of [47] rely on techniques known from aspect-based opinion mining to build a model that combines aspects with latent probabilistic user and items states in a holistic fashion. For this, they incorporate not only overall but also aspect ratings. This limits the model capacity insofar as the set of potential aspects is thereby fixed. Additionally, it requires explicit aspect ratings to be available in the first place which is commonly not the case.

Recently, in the context of the rapidly growing number of deep learning methods, researchers have exploited review data in their neural models. In [50], user and item reviews get modeled simultaneously via two Convolutional Neural Networks (CNN) with a final regression layer based on Factorization Machines [35]. Specifically, the latent states were derived from the concatenation of user and item reviews respectively. This approach was later on picked up and extended in [4]. The authors recognized that network performance can benefit not only from processing the concatenation of reviews but, simultaneously, exploiting the particular review for the current target user and target item. Another extension of [50] is presented in [5] where in addition to the overall rating the usefulness of a review is considered.

The major distinction between these approaches and ATM is that they all omit modeling aspects altogether. Additionally, although [4, 5] claim to provide explanations, these are basically complete reviews. However, using complete reviews to explain a recommendation can be counterproductive as the user will not know which parts of the review are actually important to them.

In order to highlight salient parts in a given input, e.g. texts or images, attention mechanisms have attracted considerable interest. The central idea is that the network assigns attention weights to different parts of the input and by that controls on what to focus on most. For instance, the authors of [38] combine two levels of attention, namely the local and global context of a word, to increase model transparency by hinting at important fragments of text.

But, again, they do not account for the multi-faceted nature of review texts. As already recognized by the earlier approaches [14, 34, 48], reviews usually tackle more than one topic. Therefore, an item rating has to be viewed as an overall interpretation of multiple related aspects with different weights. These weights are user-specific, that is when reading (or writing) a review, users consider different aspect to be differently important.

Yet, there exist approaches that model complex user or item profiles that are comprised of, at least latent, multi-faceted concepts. They usually act on so called memory components to incorporate structured or unstructured knowledge. Memory networks such as [40] or [11] originated from Natural Language Processing. Only recently, they have been applied to RS. The authors of [49] use a memory component in unison with attention to learn deep user representations. As opposed to our method where a memory is defined relative to aspects and sentences, the authors view a memory rather as a history of consumed items together with the respective reviews. Other works use information sources different from user reviews. The method presented in [17], for example, incorporates knowledge graph information as the foundation for a memory. Another example for this line of research is [6]. Here, the authors leverage an external memory matrix in which historical user records are stored. The main difference to ATM is that they derive memory states from mostly structured data. Our work, in contrast, aims at combining attention and memory writing and reading techniques to impose, in the first place, a structure on the input texts.

3 METHOD

The ATM architecture contains, at its core, two neural memories which can be viewed as arrays of slots for storing and thus memorizing information [11, 40]. The first memory encodes representations of sentences composed by the target user. The second one is an equivalent variant for the target item and encompasses statements by other users. Both memories are comprised of two subcomponents (see Figure 2). First, aspect-based key vectors are used to perform the addressing operation, i.e. the selection of relevant memory locations. Keys are calculated by reconstructing sentences as a weighted combination of aspect embeddings such that the memory can be read in terms of topical overlap with the query vector. Value vectors, i.e. the content encoded into memory, depict the second component and contain the encoded sentence semantics.

In order to predict the target rating, read operations extract elements from both user and item memories in a mixed-initiative fashion (see Figure 3). The user memory is first queried by an item embedding to calculate the match between user preferences and item properties. We call the result the *explicit* user state since, although latent, it is derived from sentences put into writing as an active process by the user. Since our goal is to increase network

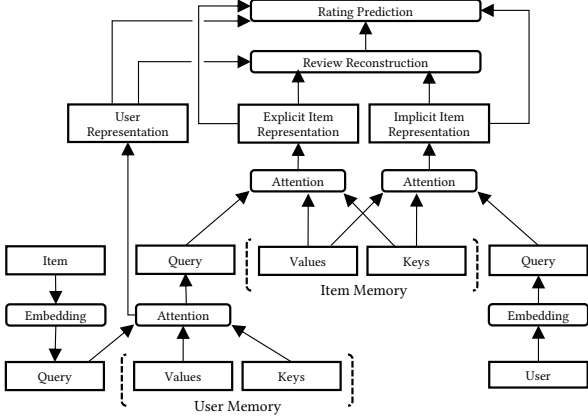


Figure 3: Schematic illustration of the ATM pipeline.

intelligibility, we translate this latent user state back into comprehensive information by computing a query for the aforementioned item memory. Aligning the item-conditioned explicit user state with the opinions of other users helps us identify those sentences that best describe how the user would likely evaluate the item.

However, certain patterns in a person’s rating behavior cannot be explained by their reviews, for instance if they are unaware of them themselves. We assume that addressing the item memory only with the explicit user state is insufficient. Therefore, we additionally train an *implicit* user representation that captures latent patterns similar to conventional collaborative filtering. This implicit state then serves as an additional query to the item memory. Both resulting vectors, explicit and implicit, can subsequently be combined to predict the target rating.

Summarized, ATM implements a mixed-initiative reading mechanism where, first, the user memory is read by the item to arrive at a latent state, and, second, the item memory is queried by explicit and implicit user representations to explicate these states.

In the following, we describe the proposed model architecture in more detail.

3.1 Preliminary

The training set consists of tuples $(u, i, r_{ui}, D_{ui}) \in \mathcal{D}$ where $r_{ui} \in \{1, 2, \dots, 5\}$ is the rating assigned by user $u \in \{1, 2, \dots, |\mathcal{U}|\}$ to item $i \in \{1, 2, \dots, |\mathcal{I}|\}$ and D_{ui} is the corresponding review. A user is additionally associated with a set of sentences $\{s\}^u$ extracted from their personal reviews. Accordingly, we find a set of sentences $\{s\}^i$ for each item. Sentences are defined as being comprised of a sequence of n embedded words $\{E[w_j]\}_{j=1, \dots, n}^s$ where $E[w_j] \in \mathbb{R}^d$ is the embedding for the j 'th word in the sentence and d is the embedding dimension. We also define a number of aspects $a \in \{1, 2, \dots, |\mathcal{A}|\}$.

3.2 User Memories

As already mentioned previously, the memory in ATM are composed of two components: Key and value vectors. We now describe the process of arriving at both for the user side, starting with the aspect-based keys. Afterwards, we explain how this established memory can be accessed by item-based queries.

We first employ an aspect extraction mechanism that learns aspect embeddings stored in an embedding matrix $A \in \mathbb{R}^{d \times |\mathcal{A}|}$ where each aspect can be interpreted by looking at the nearest words in the embedding space. The aspect embeddings can either be fixed a-priori or trained jointly with the remaining model parameters. Fixing an embedding can be done by initializing it with the embedding vectors of single words or combinations thereof.

In order to extract aspect-related information from sentences, we follow the approach proposed in [13]. We set up a sentence reconstruction pipeline that is quite similar to the concept of Autoencoders [44]. Specifically, we try to capture as much of a sentence’s semantics as possible solely by utilizing combinations of aspect representations drawn from A . Formally, a key matrix $K_u \in \mathbb{R}^{m \times d}$ is the result of stacking aspect-based reconstructions $\hat{E}[s] \in \mathbb{R}^d$ of the global sentence embeddings:

$$K_u = [\hat{E}[s_1] \quad \dots \quad \hat{E}[s_m]]^T, \quad (1)$$

where m is the number of sentences written by user u . The global sentence embeddings $E[s] \in \mathbb{R}^d$ themselves are simply computed by averaging the contained word embeddings:

$$E[s] = \frac{1}{n} \sum_{j=1}^n E[w_j] \quad (2)$$

Please note that a measure for the reconstruction quality of $E[s]$ will later be appended to our training objective. As a result, aspect embeddings can be optimized to approximate any sentence with respect to the main topics identified during training.

The general idea of approximating $E[s]$, derived from the concept of Autoencoders, is to reconstruct a sentence as a linear combination of aspect embeddings:

$$\hat{E}[s] = A \cdot p_s, \quad (3)$$

where $p_s \in [0, 1]^{|\mathcal{A}|}$ is a weight vector expressing the alignment between a given sentence and the aspects. p_s can be obtained by projecting a weighted combination of word embeddings into the aspect space:

$$p_s = \text{softmax}\left(T_{d, |\mathcal{A}|} \left(\sum_{j=1}^n b_j E[w_j] \right)\right) \quad (4)$$

where $T_{d, |\mathcal{A}|}$ is an affine transformation from d to $|\mathcal{A}|$. By inspecting the attention weights in p_s , we can estimate how strongly a sentence is aligned with each aspect.

The attention weight $b_j \in [0, 1]$ encompasses the degree to which w_j contains salient information about the sentence semantics. In other words, the attention mechanism also de-emphasizes words that do not contribute to meaning, e.g. expletives. The weights are derived subject to word embeddings as well as the global sentence embedding. Additionally, we define a matrix $Q \in \mathbb{R}^{d \times d}$ that maps between global and aspect context and is learned as part of the learning process:

$$b_j = \text{softmax}(E[w_j]^T \cdot Q \cdot E[s]) \quad (5)$$

Summarized, K_u captures a reconstruction of each sentence based on a weighted combination of aspect embeddings. In other words, the reconstruction only contains information about aspects

with the weights depicting the singular distribution of aspect importance for the given user-item combination. Hence, the keys can be viewed as a topic addressing mechanism.

The memory values on the other hand can be obtained by passing the sentence embeddings through an encoder network. While any established architecture would be feasible, we decided to apply a bidirectional Long Short-term Memory (LSTM) [16]:

$$\mathbf{h}[s] = \left[\overrightarrow{LSTM}(\{\mathbf{E}[w_j]\}_{j=1,\dots,n}^s) \oplus \overleftarrow{LSTM}(\{\mathbf{E}[w_j]\}_{j=1,\dots,n}^s) \right] \quad (6)$$

The memory value component $\mathbf{M}_u \in \mathbb{R}^{m \times d}$ is then constructed by stacking the hidden states of each sentence:

$$\mathbf{M}_u = [\mathbf{h}[s_1] \quad \dots \quad \mathbf{h}[s_m]]^T \quad (7)$$

The resulting user memory can subsequently be addressed by a given item i for which we want to make a prediction. We do this by employing another attention mechanism subject to a learned item embedding $\mathbf{E}[i] \in \mathbb{R}^d$ to identify those sentences that best match the properties of i :

$$g_{is} = \text{softmax}\left(\mathbf{T}_{d,d}(\mathbf{K}_u[s])^T \cdot \mathbf{T}_{d,d}(\mathbf{E}[i])\right), \quad (8)$$

where $\mathbf{K}_u[s]$ is the slot in memory corresponding to sentence s . The attention weight g_{is} encompasses the degree to which s is aligned with the properties of i . Next, we read from user memory according to the attention:

$$\mathbf{y}_{ui} = \sum_{j=1}^m g_{is_j} \mathbf{T}_{d,d}(\mathbf{M}_u[s_j]) \quad (9)$$

\mathbf{y}_{ui} can be interpreted as a user representation relative to target item and observed aspects.

3.3 Item Memories

The item side of ATM is architecturally equivalent to the user side. Therefore, the item memory \mathbf{M}_i can be computed in accordance to the process described for the user side. Please note that parameter weights can be shared between item and user side. The difference between both model components is limited to the reading operation. Instead of only having access to a single latent query embedding $\mathbf{E}[u] \in \mathbb{R}^d$, ATM also refers to the already processed explicit information about the user. Unfortunately, the value vectors in \mathbf{M}_u were encoded independent of any notion about aspects and, therefore, do not suffice as a query for the item memory. As an alternative, however, we can simply calculate an explicit query vector $\mathbf{q}_{ui} \in \mathbb{R}^d$ by applying the attention weights g_{is} to the key vectors \mathbf{K}_u :

$$\mathbf{q}_{ui} = \sum_{j=1}^m g_{is_j} \mathbf{T}_{d,d}(\mathbf{K}_u[s_j]) \quad (10)$$

Since the ultimate goal is to turn latent knowledge transparent, we need to associate both user queries, explicit and implicit, with sentences by other users about the target item i . To retrieve explicit, $\mathbf{z}_{ui}^e \in \mathbb{R}^d$, and implicit representations, $\mathbf{z}_{ui}^i \in \mathbb{R}^d$, respectively, we attentively query the item memory twice:

$$\begin{aligned} g_{us}^e &= \text{softmax}\left(\mathbf{T}_{d,d}(\mathbf{K}_i[s])^T \cdot \mathbf{T}_{d,d}(\mathbf{q}_{ui})\right) \\ g_{us}^i &= \text{softmax}\left(\mathbf{T}_{d,d}(\mathbf{K}_i[s])^T \cdot \mathbf{T}_{d,d}(\mathbf{E}[u])\right) \end{aligned} \quad (11)$$

The final representations can then be calculated by:

$$\begin{aligned} \mathbf{z}_{ui}^e &= \sum_{j=1}^m g_{us_j}^e \mathbf{T}_{d,d}(\mathbf{M}_i[s_j]) \\ \mathbf{z}_{ui}^i &= \sum_{j=1}^m g_{us_j}^i \mathbf{T}_{d,d}(\mathbf{M}_i[s_j]) \end{aligned} \quad (12)$$

3.4 Review Reconstruction & Rating Prediction

The resulting states are utilized to approximate two objectives: First, it has been shown repeatedly [4, 50] that much of the predictive value for a rating r_{ui} stems from the target review D_{ui} . Therefore, we feed both \mathbf{z}_{ui}^e and \mathbf{z}_{ui}^i as well as \mathbf{y}_{ui} into a review reconstruction layer that tries to predict the actual embedding of the target review. This comes with the additional benefit that the network will eventually learn to assign large weights to sentences in the explicit item memory that follow a linguistic style similar to the target user's. This is important because meaning is highly subjective and hard to extract unambiguously from written text. Human language is filled with subtleties and idiosyncrasies. For instance, one person might want to express a distinctly positive experience when stating that a movie was *nice* while for another it is just a polite way of saying that they found it underwhelming. It can be difficult or impossible, even for humans, to relate an opinion expressed by an anonymous stranger with their own views. By augmenting the architecture with a review reconstruction component, we try to strike the right note when selecting sentences.

Formally, the concatenated states are fed through a feed-forward network to arrive at a latent reconstruction of the embedded target review $\mathbf{E}[D_{ui}]$:

$$\hat{\mathbf{E}}[D_{ui}] = \mathbf{T}_{d,d}\left(\mathbf{T}_{3d,d}^{\text{prelu}}(\mathbf{y}_{ui} \oplus \mathbf{z}_{ui}^e \oplus \mathbf{z}_{ui}^i)\right), \quad (13)$$

where $\mathbf{T}_{2d,d}^{\text{prelu}}$ is an affine transformation followed by a PReLU activation function [12]. The reconstructed review embedding in conjunction with the remaining state vectors states can then be used in a final layer to predict the rating score:

$$\hat{r}_{ui} = \mathbf{T}_{4d,1}(\mathbf{y}_{ui} \oplus \mathbf{z}_{ui}^e \oplus \mathbf{z}_{ui}^i \oplus \hat{\mathbf{E}}[D_{ui}]) \quad (14)$$

3.5 Training Objective

In order to optimize the proposed model, we design a multi-criteria loss function which we will explain component-wise. The first one is a conventional rating prediction target, i.e. mean squared error:

$$J^r(\theta) = \sum_{(u,i)} (r_{ui} - \hat{r}_{ui})^2 \quad (15)$$

Secondly, we consider the reconstruction loss of the target review D_{ui} . For this, we first need to embed the actual review into latent space. In our case, this is done by simply averaging over the contained word embeddings although more sophisticated variants such as Generative Adversarial Networks [10] are applicable, too. To optimize reconstruction quality, we utilize a hinge loss maximizing the similarity between reconstruction and target at the same time as minimizing the similarity between reconstruction and negative samples $D_{u',i'}$:

Table 1: Data set summary.

Dataset	#items	#users	#reviews	density
Yelp	156 863	48 894	2 268 664	$2.958 \cdot 10^{-4}$
Movies	123 952	50 052	1 691 457	$2.726 \cdot 10^{-4}$
Kindle	68 222	61 932	978 820	$2.317 \cdot 10^{-4}$

$$J^D(\theta) = \sum_{(u,i) \in \mathcal{D}} \sum_{(u',i') \in \mathcal{S}} \max(0, 1 - \hat{\mathbf{E}}[D_{ui}]^T \cdot \mathbf{E}[D_{ui}] + \hat{\mathbf{E}}[D_{ui}]^T \cdot \mathbf{E}[D_{u'i'}]), \tag{16}$$

where $\mathcal{S} \subseteq \mathcal{D} \setminus (u, i)$. Equivalently, we can formulate a loss $J^S(\theta)$ for the aspect-based sentence reconstruction between $\mathbf{E}[s]$ and $\hat{\mathbf{E}}[s]$.

Finally, we also include a regularization term to prevent learned aspect from becoming too similar over time:

$$U(\theta) = \|\mathbf{A} \cdot \mathbf{A}^T - \mathbf{I}\|, \tag{17}$$

where \mathbf{I} is the identity matrix. Without introducing this regularization, the learned vector representations for aspects may suffer from redundancy problems. Appending $U(\theta)$ to the objective function encourages the diversity of the resulting embeddings.

The final objective function, hence, is obtained by:

$$L(\theta) = J^r(\theta) + \lambda_D J^D(\theta) + \lambda_S J^S(\theta) + \lambda_U U(\theta) \tag{18}$$

with λ being hyperparameters in the range $[0, 1]$.

4 EVALUATION

The evaluation is subdivided into three sections: In the first one, we compare the offline performance of ATM against two baselines on real-world datasets. The second part is concerned with qualitative analyses of specific model properties. Concretely, in a case study we inspect the distribution of attention weights among sentences on the user- as well as on the item-side to ascertain how ATM utilizes its internal information base for the personalization of recommendations and explanations. In the third section, we present the results of an online user study we have conducted to find out what constitutes a good textual explanation in a decision-making task. Specifically, we compare the performance of ATM explanations against a state-of-the-art review-retrieval method and show that, while both approaches yield adequate results, aspect-based explanations are overall preferred by users.

4.1 Offline Experiments

We have conducted experiments on three real-world datasets to demonstrate the effectiveness of ATM by comparing it to state-of-the-art RS. In the following, we present datasets, describe our experimental procedure, and introduce the selected baselines for comparison. Then, we evaluate and discuss performance.

4.1.1 Datasets. The *Yelp*¹ dataset is a large-scale dataset introduced in the context of the Yelp challenge. Since aspects only relate to specific domains, we filtered out all reviews for businesses not associated with the category *Restaurants*. *Kindle* is one category of the Amazon review dataset² containing reviews of e-books purchased from the Kindle store. *Movies* is another of the Amazon categories

¹<https://www.yelp.com/dataset/challenge>

²<http://jmcauley.ucsd.edu/data/amazon/>

with movie and TV reviews. All datasets contain user reviews associated with a 5-star overall rating. The considered datasets are summarized in Table 1.

4.1.2 Procedure and Settings. In our experiments, we adopted the well-known Mean Squared Error (MSE) metric to evaluate recommender performance. We have decided to use this metric in order to maintain some degree of comparability to related works that also employed the same metric.

Preprocessing and Embeddings. Reviews were first passed through a Stanford Core NLP Tokenizer [27] to obtain tokens which were then lowercased. Sentences were separated subject to the tokenizer result. Contractions were expanded and stopwords and punctuation were combined into a single token. We set a maximum number of 30 words per sentence with a total of 150 sentences per user and item. Shorter sentences were padded accordingly. We used pretrained *fastText* embeddings [18] with dimensionality 300 for word embeddings.

Initialization. We trained a variant of our proposed model with 10 aspects initialized with random values drawn from a Glorot uniform distribution [9]. Please refer to the evaluation in [13] to verify that the process of learning aspect embeddings described in Section 3 yields meaningful results. Random initialization of the remaining parameters was also done with respect to a Glorot uniform distribution. Concerning the model-specific hyperparameters, we set $\lambda_D = 0.6$, $\lambda_S = 1.0$ and $\lambda_U = 1.0$. All of these values were selected via grid-search-like optimization.

Training and Testing. Optimization was achieved using Adam [20] and a learning rate of 0.001. We randomly split the data into training (80%), validation (10%), and test set (10%). The maximum number of epochs was set to 10. After training for one epoch with a batch size of 32, we calculated MSE on validation and test set. Similar to [4], we report the results of the test set where the results of the validation set was lowest. All algorithms were implemented with Python using PyTorch [31].

4.1.3 Baselines. We compare our method against established recommendation models:

- **Matrix Factorization (MF)** [23] is one of the most popular collaborative filtering techniques.
- **Neural Attentional Rating Regression (NARRE)** [5] is a CNN-based model that consists of two parallel attentive neural networks coupled by a final recommendation layer. The first network processes reviews of a target user in an attentive manner to derive a latent state. The second does the corresponding operation for the item side. Since we were mainly interested in comparing review-retrieval approaches with our aspect-based variant, we view NARRE, known to produce state-of-the-art recommendations, as a pars pro toto for the whole line of research.

The parameter values for NARRE were assigned subject to the evaluation in the original paper. The textual data made available to NARRE was chosen to match the total amount available to ATM.

4.1.4 Offline Performance. The results for the rating prediction task for ATM and the baselines are given in Table 2. As was to be

Table 2: MSE for all baselines as well as the proposed ATM.

	Yelp	Kindle	Movies
Baselines			
MF	1.379	0.739	1.488
NARRE	1.102	0.519	0.922
ATM	1.085	0.454	0.847

expected, the two review-based approaches perform better than the conventional collaborative filtering model.

Furthermore, our proposed model outperforms NARRE on all considered datasets. One explanation for this is that its mixed-initiative approach allows ATM to more selectively distribute attention. While the integration between target user and item only happens in the later layers in NARRE, user and item side are interwoven in ATM right from the beginning. Additionally, breaking reviews down into sentences allows the model to distribute its attention on smaller semantic units. Finally, the integration of the review reconstruction pipeline strengthens the overall training signal.

4.2 Qualitative Analysis

In this section, we take a closer look at the attention dynamics in the network with respect to one user, who we call *Alice* for convenience, and the movie *Gravity*. The model was trained with three fixed aspects: *Story & Setting*, *Actors & Characters*, and *Cinematography*. In Figure 4, two sentences are depicted for each aspect. We have selected the sentences with the highest scores obtained when multiplying aspect attention p_s and the individual sentence attention values g_{is} , g_{us}^e , or g_{us}^i respectively.

Please note that we, on purpose, kept Alice’s review D_{ui} for *Gravity* available in the user memory. Although obviously impossible in real-life applications, the intention behind this was as follows: If the network would assign the largest attention weights to the sentences contained in D_{ui} , this would verify that the network is in general capable of determining the most informative memory slots for the current item. And indeed, as Figure 4 indicates, this is true for five out of six sentences.

The average attention distribution over sentences in the user memory reveals that, by far, the strongest focus (0.85) is assigned to the *Story & Setting* aspect. In other words, Alice generally seems to be concerned with storytelling when writing reviews. In this case, she mainly complains about the dialog being underwhelming, sometimes even cringeworthy. This opinion overshadows the otherwise positive impression of acting and cinematography.

The attention weights on the item side are more evenly distributed indicating that users, in general, discuss *Gravity* from various perspectives with a slight tendency towards *Actors & Characters* (0.45). Interestingly, in case of the explicit state, the tone for *Story & Setting* is also quite negative. However, instead of targeting the movie’s dialogs, the extracted sentences are more concerned with the plot itself. Such a negative connotation is found for *Actors & Characters* as well. In this case, however, there is an obvious discrepancy between the network’s inference, i.e. disliking the acting, and the actual user impression, i.e. being impressed by the cast. This finding stresses that any automatically generated explanation for a recommendation is crucially subject to uncertainty. Supposedly,

this uncertainty will be greater the more fine-grained the analysis becomes.

In case of the implicit representation, we can observe the exact opposite for both *Story & Setting* as well as *Actors & Characters*. It is particularly interesting that the assessment of the actors’ performance is much closer to the expressed opinion in the actual review. On the other hand, the extracted sentences for *Story & Setting* not only miss Alice’s focus on dialog, but also falsely assume an overall positive sentiment for the aspect.

While some user preferences can indeed be extracted from old reviews, see aspect *Cinematography*, every new movie is ultimately a novel experience which may underly preferential shifts. Some preferences might have been unknown to Alice until watching *Gravity* or will remain so forever. On the other hand, implicit representations suffer more severely from sparse data making it extremely hard to identify behavioral patterns properly. The bottom line here is that ATM crucially relies on both explicit and implicit states to derive a comprehensive representation.

4.3 User Study

We hypothesized that generating and presenting explanations in an aspect-based fashion would improve the subjective assessment of explanation quality compared to the retrieval of complete reviews. We further assumed that the distillation of relevant sentences and grouping them with respect to aspects would increase the overall textual quality of explanations.

In particular, generating aspect-level explanations may help improve linguistic coherence by removing redundancies, repetitions and filler phrases from the explanations. Aspect-based explanations not only impact the textual content itself, but also the way how explanations can be presented to the user. Instead of being restricted to complete reviews, ATM can group sentences by aspects which may result in a more orderly visual presentation of information. This, however, may come with the unwanted side effect of unhinging sentences from their original context which may negatively influence the argumentative flow. Summarized, we hypothesized that the overall assessment of explanation quality can in fact be subdivided into three key factors: *Content adequacy*, *linguistic adequacy* as well as *presentation adequacy*.

4.3.1 Method. To study the aforementioned hypotheses, we conducted an online user study with Amazon Mechanical Turk requiring participants to be located in the US and to have an approval rate greater than 95%. We assigned them to two conditions in counter-balanced order in a between-subject design. In both conditions, participants were presented with five randomly sampled movies from the 100 most popular movies (estimated by the total number of reviews) from the Amazon movie dataset.

Each movie was depicted in terms of the following variables: *Title*, *runtime*, *director*, *actors* as well as a movie *poster*. Additionally, users received an explanation generated subject to the condition they were assigned to. Users in the first condition were shown aspect-based explanations generated by ATM whereas users in the second condition received the most helpful reviews yielded by NARRE. For ATM, we extracted the three most strongly associated sentences for each of the following four manually selected aspects:

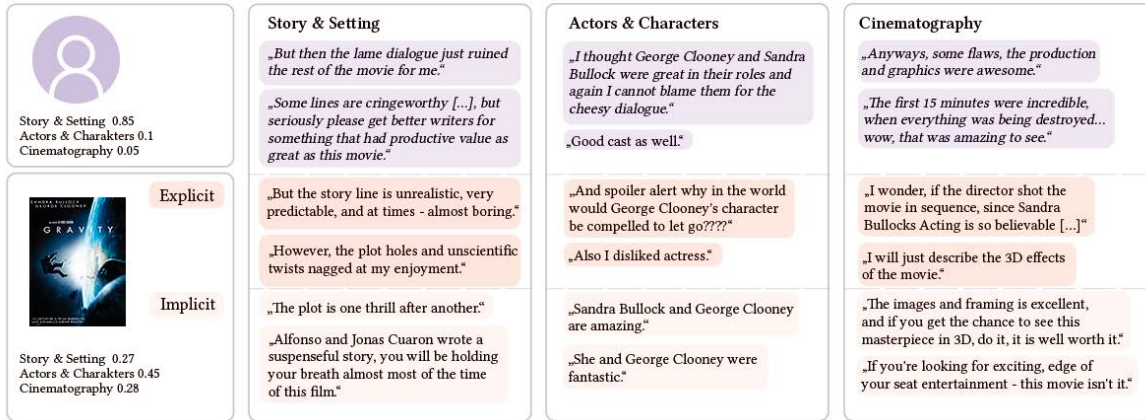


Figure 4: Case study for one user, the movie Gravity and three aspects. The shown sentences are the ones with the highest activation for the respective aspects. Sentences in italic are drawn from this user’s review of Gravity.

Table 3: Constructs assessed during the user study. Factor loadings are depicted for constructs with multiple item.

Construct	Items
Explanation Quality	I liked the explanations provided by Streamflix. (0.831)
	Streamflix is providing good explanations. (0.854)
Content Adequacy	The explanations were relevant. (0.731)
	The explanations provided for the movies are sufficient for me to make a decision. (0.786)
Presentation Adequacy	The explanation content was sufficient to get a good overview of the movie. (0.837)
	The explanations were presented adequately.
Linguistic Adequacy	The presentation of ideas was not orderly. (0.890)
	There were abrupt shifts in the explanations. (0.804)
	The explanations were not direct in their approach to the subject. (0.850)

Story & Setting, Characters & Actors, Cinematography, Uniqueness, i.e. what reviewers think is special about this movie.

In order to ensure comparability, we constrained the overall number of sentences to be roughly the same between both conditions. The average number of sentences per review in the dataset is 15 such that most movies were described by one or two reviews in the NARRE condition. For both conditions, we manually removed those reviews and sentences that mostly dealt with topics irrelevant to the task, e.g. discussions of the quality of the DVD.

We, on purpose, withheld a movie synopsis as we were interested in whether sufficient information could be extracted from the provided explanations alone.

Table 4: Mean Values and Standard Deviations for dependent variables which were assessed on a 5-point Likert scale.

Variable	ATM		NARRE	
	M	SD	M	SD
Explanation quality	3.94	0.83	3.61	0.96
Content adequacy	3.90	0.98	3.64	1.11
Presentation adequacy	3.94	1.01	3.39	1.10
Linguistic adequacy	3.24	1.32	3.0	1.05

Please note that this study was mainly designed to investigate general differences between aspect-based explanations on sentence-level and review-level explanations. Hence, we omitted the personalization component in both approaches as this would require participants to actually write reviews which was out of the scope for this study setup. Therefore, both models were trained with the user side held constant updating only item-related parameters.

Procedure: First, participants were instructed that they were about to interact with a novel movie streaming service. In order to elicit their preferences, five randomly selected movies would be displayed and their task was to select one of them that fits their preferences best. The interaction took place in a small web application which allowed participants to access available meta-data and explanations for each movie (Figure 1). Afterwards, participants were asked to rate movies and explanations as well as provide an overall impression of system-related aspects by filling in questionnaires.

Questionnaire: For composing the questionnaire, we partially relied on established RS evaluation instruments. In order to estimate *explanation quality* and information sufficiency (operationalized as *content adequacy*), we used constructs from [21]. We also generated items ourselves that we thought would serve the evaluation of the hypothesized factors of explanation quality. See Table 3 for a summary of the items and constructs used for the study. All items were assessed on a 1–5 Likert-scale.

4.3.2 Results. Descriptive results of our study can be found in Table 4. They are split subject to the experimental *condition*. In order to unravel by which dimensions our experimental conditions influence the quality of explanations and how these dimensions relate to each other, we hypothesized a structural model (see Figure 5) that we will describe in the following.

Based on the number of latent constructs and observed variables we estimated the lower-bound for the sample size. With the probability level set to $\alpha = 0.05$ and a desired statistical power level of 0.8, the sample is required to be comprised of 136 observations to, at least, detect medium effects (0.3) [7, 46].

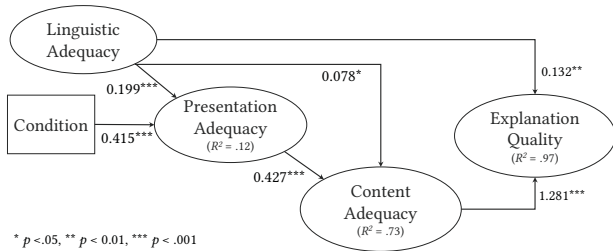


Figure 5: Structural Equation Model comparing the influence of ATM vs. Narre. The edges show standardized parameter weights and the amount of explained variance for endogenous variables is displayed inside the nodes.

We were interested in identifying whether the interaction with two different types of explanations led to differences in the assessment of explanation quality. *Condition* was defined as an exogenous categorical variable. We hypothesized that *condition* had an impact on *explanation quality* that is mediated by *linguistic*, *presentation*, and *content adequacy*. We further assumed that *presentation adequacy* depends on linguistic aspects while both partially influence content. Structural equation modeling was applied to trace causal paths that predict *explanation quality*. For this, we utilized the *R* package *lavaan*, version 0.6-5 [36].

We conducted outlier detection, a test for normality, and the selection of an appropriate estimator as preparation steps. Outlier detection based on Cook’s distance revealed eight rows which were subsequently dropped. Additionally, we excluded participants that finished the interaction in under five minutes, leaving us with a final sample size of 148 participants (65 female), average age 38.5 ($SD = 11.2$). Shapiro’s test for normality indicated that several variables of interest significantly deviated from normal distributions. As a result, we ran the analysis with the MLM estimator that allows for robust standard errors and scaled test statistics [8].

Our hypothesized model appears to be a good fit for the data ($CLI = .989$, $TLI = 0.981$, $RMSEA = 0.042$). For the sake of clarity, we report significant direct effects successively from left to right. Along these paths, we trace back mediated influences from *condition* on the endogenous variables. Constructs assessed with more than one item showed sufficient factors loadings (see Table 4).

Direct Effects. The positive direct effect from *condition* onto *presentation adequacy* (see Figure 5) suggests that structuring explanations into aspect groups is meaningful. *Presentation adequacy* is additionally influenced by *linguistic adequacy*. Linguistic weaknesses like redundancies, abrupt shifts or a verbose writing style negatively impact presentation. Based on this assumption, it can be deduced that the direct effect of *condition* cannot be traced back to differences in the linguistic dimension between both algorithms, but exclusively with respect to the way how information is depicted.

We could not find a significant direct effect from *condition* on *linguistic adequacy* indicating that, although complete reviews tend to incorporate redundancies and irrelevant sections, aspect-based explanations still suffer from comparable drawbacks.

Concerning *content adequacy*, we can observe a direct effect from both *linguistic* and *presentation adequacy*. Hence, the assessment of

content itself appears to depend on the way the text is formulated and presented.

As hypothesized, *linguistic* and *content adequacy* both significantly influence the overall assessment of *explanation quality*. Although no direct influence from *presentation adequacy* is present, closer inspection of the structural model indicates that there exists, in fact, a significant effect on *explanation quality*. However, it turns out that this effect is fully mediated by *content adequacy*. In the following, we describe those mediating effects in detail to shed light on the complex interactions between the observed constructs as well as our experimental condition.

Mediated Effects. As just described, the total effect from *presentation adequacy* on *explanation quality* is mediated by the route [*presentation adequacy* → *content adequacy* → *explanation quality*] with a standardized parameter weight of 0.55 ($p < .001$). In a similar manner, the total effect from *linguistic adequacy* on *explanation quality* can be calculated by combining the direct effect taken from the model with the indirect effects of the two routes [*linguistic adequacy* → *content adequacy* → *explanation quality*] and [*linguistic adequacy* → *presentation adequacy* → *content adequacy* → *explanation quality*] resulting in a total effect of 0.34 ($p < .001$).

No significant direct effect from *condition* on any other construct than *presentation adequacy* can be identified. However, we can trace back mediated effects on *content adequacy* again via [*condition* → *presentation adequacy* → *content adequacy*]. By doing so, a standardized weight of 0.18 ($p = 0.019$) can be derived. Therefore, the benefit of aspect-based explanations does not stem from a higher content quality of the extracted sentences but from the way these sentences are presented. The mediated effect [*condition* → *presentation adequacy* → *content adequacy* → *explanation quality*] yields a standardized weight of 0.23 ($p = 0.01$).

4.3.3 Discussion. Inspection of the structural model hints at a pivotal role for the hypothesized factors of explanation quality. We found that the total effect on *explanation quality* was significant for *content adequacy*, *presentation adequacy* as well as *linguistic adequacy*. Combined, they explain a very high amount of 97% of the variance in *explanation quality* making it safe to assume that these three factors are indeed crucial indicators for the quality of textual explanations. This finding stresses that the evaluation of explanations in RS is by no means a one-dimensional problem. Instead, there seem to exist multiple interacting factors contributing simultaneously to whether a person perceives an explanation as helpful or not. RS research should therefore shift away from evaluating *explanation quality* by means of offline metrics. Instead, a user-centric perspective enables us to view the process of constructing explanations as a multi-faceted task that includes the identification of relevant topics or aspects as well as the optimization of linguistic as well as user experience parameters.

Concretely, we found that *content adequacy* is, by far, the most influential predictor for *explanation quality*. It seems natural to assume that the content itself should have the strongest contribution. However, the adequacy of content cannot be looked at in isolation. If the explanation is too wordy or suffers from abrupt shifts or redundancies, i.e. weaknesses in *linguistic adequacy*, the perception of content quality also decreases. The effect is, in fact, two-fold: First, linguistic cohesion directly affects content in the sense that

cognitive comprehension performance will significantly improve if the explanation follows a clear and correct syntax, vocabulary, and argumentative flow [39]. Second, *linguistic adequacy* also has a significant effect on *presentation adequacy* since text parameters such as length are directly linked to visual depiction of information. *Presentation adequacy*, in turn, is directly connected to reading comprehension [19]. From a psychological perspective, *linguistic* and *presentation adequacy* can be viewed as factors that determine the difficulty of the textual material as described by cognitive load theory [41]. Reading and processing explanations can be considered as a learning task where intellectual activity and schema acquisition are the primary mechanisms. Subject to cognitive load theory, structuring of information in order to reduce difficulty is extremely important, especially in the case of RS where people often have to process information concerning multiple items at the same time. That is, the intrinsic cognitive load of the decision-making task can be extremely high even if an RS reduces the number of valid alternatives. Pros and cons still have to be considered and explanations, if present, still need to be processed. It follows that extraneous factors such as arrangement and wording of explanations need to inflict as little cognitive load on the user as possible.

Transferring these findings to the evaluation of our experimental condition, we observe that ATM achieves higher level of explanation quality compared to NARRE (see Table 4). People are more confident when making a decision based on the explanations provided by ATM and found them to be more relevant. Interestingly, this finding cannot be explained by an enhanced *content adequacy*. Although there exists a significant total effect from condition on *content adequacy*, by distinguishing between direct and indirect effects, we were able to detect a systematic effect that is fully mediated by *presentation adequacy*. In other words, the improved perceived quality of explanations can be traced back solely to the way ATM arranges the explanations. Apparently, aspect-based explanations ease the comprehension of textual information by ordering them in a way meaningful to the user which is in line with cognitive load theory. By sorting sentences into groups of aspects, the RS may partially adopt the task of activating respective cognitive schemes resulting in reduced cognitive load.

On the same note, ATM and NARRE can both benefit from improving the linguistic adequacy of their explanations. Generally, explanations should not be too wordy, present their ideas in an orderly fashion as well as prevent redundancies and abrupt shifts. Concerning all these aspects, both approaches still have room for improvement. Therefore, it is of importance that explainable algorithms are not only designed to find relevant information per se, but also to incorporate means of how to formulate them adequately. Unfortunately, the presented study yields no further statistical insights into what factors account for differences in *linguistic adequacy*. However, users frequently indicated in the NARRE condition that, e.g., “some of them were short and some of them were way too long and gave too many details” which points at one particular drawback of retrieving full reviews: A general lack of control over the explanation, especially its length.

Concluding, ATM and NARRE achieve a reasonable level of *explanation quality* indicating that both provide a sufficient amount of information to improve the decision-making. But, still, the question remains what, in particular, constitutes a good explanation.

One drawback of our study is the fact that we completely left out the personalization component which is usually at the core of any RS. For instance, someone wrote concerning the ATM condition: “The textual explanations for these films were extremely satisfying. However, I think [the algorithm] should have insisted more on the sentimental and artistic side of these films.” This comment hints that the decision of restricting the aspects to a fixed set may lead to suboptimal results by missing subtle personal preferences. Therefore, our future research should directly target at evaluating explanations that are extracted on an individual level.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed ATM, an approach to memorize user opinions on relevant item aspects found in raw review texts to derive multi-faceted user and item representations. We have shown that representing knowledge about multiple aspects in combination with external memories leads to accurate recommendations. Offline experiments indicate that ATM outperforms other review-based models by at least a slight margin. A qualitative analysis of attention weights allowed us to look deeper into how the network behaves in practice. Specifically we could verify that the activations in the respective memories can easily be interpreted which is a prerequisite of any explainable system that depends on an attention mechanism. In the context of a user study we could also show that ATM produces explanations of high quality.

However, the evaluation results also indicate some limitations of our work. Generally, our solution is no whitebox model [32] as some important operations remain undisclosed. For instance, although the network transparently conveys how attention is distributed, it is unclear which semantics are actually derived and encoded into the latent representations (see Section 4.2).

One possibility to further increase intelligibility would be the incorporation of sophisticated linguistic preprocessing. *Linguistic adequacy* plays a significant role in generating high quality textual explanations (see Section 4.3). In fact, however, our approach currently disregards any language structure beyond discriminating sentences. This deficit becomes apparent by observing that *linguistic adequacy* was, on average, rated lowest of all three factors of *explanation quality*. Identifying, for example, presence, polarity, and quality of arguments via argument mining [29] or stance detection [1], may help substantially improve the understanding of an author’s motives and reasoning. Furthermore, extracting information on sentence-level leads to explanations being detached from their original context. This effect is also mentioned in user comments: “I thought they were somewhat helpful, but felt a little incomplete. It was hard to draw an overall picture because the summaries were so short.” In-depth insights into entailment and semantic cohesion are other important directions for our future research.

6 ACKNOWLEDGMENTS

This work has been partially supported by the Deutsche Forschungsgemeinschaft (DFG) within the project “Argument-Based Decision Support for Recommender Systems (ASSURE)” that is part of the Priority Program “Robust Argumentation Machines (RATIO)” (SPP-1999).

REFERENCES

- [1] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in on-line debate. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*. Association for Computational Linguistics, 1–9.
- [2] Georgios Askalidis and Edward C Malthouse. 2016. The value of online customer reviews. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM.
- [3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [4] Rose Catherine and William Cohen. 2017. Transnets: Learning to transform for recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 288–296.
- [5] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1583–1592.
- [6] Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiayi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2018. Sequential recommendation with user memory networks. In *Proceedings of the eleventh ACM international conference on web search and data mining*. ACM, 108–116.
- [7] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.
- [8] David A Freedman. 2006. On the so-called “Huber sandwich estimator” and “robust standard errors”. *The American Statistician* 60, 4 (2006), 299–302.
- [9] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 249–256.
- [10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.
- [11] Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural Turing machines. *arXiv preprint arXiv:1410.5401* (2014).
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.
- [13] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 388–397.
- [14] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1661–1670.
- [15] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.
- [16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [17] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y Chang. 2018. Improving sequential recommendation with knowledge-enhanced memory networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 505–514.
- [18] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016).
- [19] Marcel A Just, Patricia A Carpenter, and Jacqueline D Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of experimental psychology: General* 111, 2 (1982), 228.
- [20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [21] Bart P Knijnenburg, Martijn C Willemsen, and Alfred Kobsa. 2011. A pragmatic procedure to support the user-centric evaluation of recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 321–324.
- [22] Joseph A Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User modeling and user-adapted interaction* 22, 1-2 (2012), 101–123.
- [23] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.
- [24] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *CHI Conference on Human Factors in Computing Systems Proceedings*. ACM, New York, NY, USA, to appear. <https://doi.org/10.1145/3290605.3300717>
- [25] Béatrice Lamche, Ugur Adigüzel, and Wolfgang Würndl. 2014. Interactive explanations in mobile shopping recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*. 14.
- [26] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).
- [27] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.
- [28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
- [29] Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*. ACM, 98–107.
- [30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.
- [31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).
- [32] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2017. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*. ACM, 1–18.
- [33] Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating recommender systems from the user’s perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 317–355.
- [34] Zhaochun Ren, Shangsong Liang, Piji Li, Shuaiqiang Wang, and Maarten de Rijke. 2017. Social collaborative viewpoint regression with explainable recommendations. In *Proceedings of the tenth ACM international conference on web search and data mining*. ACM, 485–494.
- [35] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International Conference on Data Mining*. IEEE, 995–1000.
- [36] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012), 1–36. <http://www.jstatsoft.org/v48/i02/>
- [37] Badrul Munir Sarwar, George Karypis, Joseph A Konstan, John Riedl, and others. 2001. Item-based collaborative filtering recommendation algorithms. *Www* 1 (2001), 285–295.
- [38] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 297–305.
- [39] Mercedes Spencer, Allison F Gilmour, Amanda C Miller, Angela M Emerson, Neena M Saha, and Laurie E Cutting. 2019. Understanding the influence of text complexity and question type on reading outcomes. *Reading and writing* 32, 3 (2019), 603–637.
- [40] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, and others. 2015. End-to-end memory networks. In *Advances in neural information processing systems*. 2440–2448.
- [41] John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction* 4, 4 (1994), 295–312.
- [42] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.
- [43] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent user interfaces*. ACM, 47–56.
- [44] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. ACM, 1096–1103.
- [45] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable recommendation via multi-task learning in opinionated text data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 165–174.
- [46] J Christopher Westland. 2010. Lower bounds on sample size in structural equation modeling. *Electronic commerce research and applications* 9, 6 (2010), 476–487.
- [47] Yao Wu and Martin Ester. 2015. Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 199–208.
- [48] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 83–92.
- [49] Lei Zheng, Chun-Ta Lu, Lifang He, Sihong Xie, Vahid Noroozi, He Huang, and Philip S Yu. 2018. Mars: Memory attention-aware recommender system. *arXiv preprint arXiv:1805.07037* (2018).
- [50] Lei Zheng, Vahid Noroozi, and Philip S Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 425–434.

The following article is reused from:

Donkers, T., & Ziegler, J. (2020). Leveraging Arguments in User Reviews for Generating and Explaining Recommendations. *Datenbank-Spektrum*, 20(2), 181–187.
<https://doi.org/10.1007/s13222-020-00350-y>



Leveraging Arguments in User Reviews for Generating and Explaining Recommendations

Tim Donkers¹ · Jürgen Ziegler¹

Received: 20 February 2020 / Accepted: 10 June 2020 / Published online: 1 July 2020
© The Author(s) 2020

Abstract

Review texts constitute a valuable source for making system-generated recommendations both more accurate and more transparent. Reviews typically contain statements providing argumentative support for a given item rating that can be exploited to explain the recommended items in a personalized manner. We propose a novel method called Aspect-based Transparent Memories (ATM) to model user preferences with respect to relevant aspects and compare them to item properties to predict ratings, and, by the same mechanism, explain why an item is recommended. The ATM architecture consists of two neural memories that can be viewed as arrays of slots for storing information about users and items. The first memory component encodes representations of sentences composed by the target user while the second holds an equivalent representation for the target item based on statements of other users. An offline evaluation was performed with three datasets, showing advantages over two baselines, the well-established Matrix Factorization technique and a recent competitive representative of neural attentional recommender techniques.

Keywords Recommender Systems · Explanations · Memory Networks

1 Introduction

Deciding which news articles to read, which product to buy, or which hotel to book has become an increasingly difficult task for web users due to the sheer amount of options available. In recent years, recommender systems (RS) have become well-established tools for alleviating the user's search and decision-making in such applications [25]. A recommendation issued by such a system can be considered a specific form of a claim, namely that the user will find the recommended item useful or pleasing. In contrast to classic argumentation theory, a recommendation claims neither general nor exclusive validity but is often personalized and may depend on local, temporal or other contextual factors. Recommendations typically do not aim at influencing a person's long-term beliefs, rather, they aim at supporting users in their decision-making in a specific interactive context such as an online shop, therefore also involving a strongly persuasive component.

Conventional recommender systems mostly function as black boxes and do not provide the user with explanations why a recommendation is given. This problem has stimulated considerable research into transparency and explainability of recommendations [30]. Explaining a recommendation aims at providing supportive evidence for the claimed suitability of the recommended item. The relation between a recommendation and its explanation can therefore be considered a specific form of argumentation, although very little research has thus far investigated explainability from this perspective [4, 22]. Since current RS mostly rely on quantitative approaches, explanations can usually not be derived from explicit system inferences but are mainly based on statistical concepts, depending on the recommendation approach taken. In the popular approach of Collaborative Filtering, for example, recommendations as well as explanations are based on item ratings given by users with similar preferences, following a form of *argumentum ad populum* scheme [7]. Content-based RS derive their recommendations from the similarity between a user's preferences and the (objective) properties of an object, enabling feature-based explanations (for a comparison of methods, see [6]), while hybrid systems apply a mixture of methods.

✉ Jürgen Ziegler
juergen.ziegler@uni-due.de

¹ Universität Duisburg-Essen, Forsthausweg 2, Duisburg, Germany

In addition to these basic approaches, user-generated content has increasingly been used for generating as well as for explaining recommendations, exploiting, for example, user-provided tags [19] or textual user reviews [32]. While textual feedback from other users has been shown to support and influence decision making [1], extracting item-related aspects and sentiments from reviews is still a challenging task. It is essential, however, for producing review-based explanations. A further challenge relates to the (plausible) assumption that the relevance of a particular review for the user's decision-making is both dependent on the user's own preferences and on the convincingness of the argumentation in the review. Determining the quality and convincingness of arguments in reviews, however, is a largely open research problem.

The ASSURE project, carried out in cooperation between the Interactive Systems Group (Prof. Jürgen Ziegler) and the Language Technology Lab (Prof. Torsten Zesch of the University of Duisburg-Essen, aims at leveraging review content for improving the accuracy of personalized recommendations as well as the quality of explanations, in particular by providing argumentative explanations. In this paper, we address the problem of explaining recommendations based on aspects extracted from reviews and present a novel neural architecture for modeling both user preferences and item-related aspects.

2 Goals & Challenges

Although explanations in RS can be discussed with respect to argumentation theory, this link has rarely been established in research. The most common forms of explanations, i.e. collaborative and feature-based, rely on statistical correlations found in the data and, thus, depict an abstract form of argumentation. While explanations based on textual feedback by other users more closely resemble how humans communicate with each other and usually provide deeper insights into an item's properties, principles of argumentation theory are generally not considered during their generation process. One reason for this is that manual as well as automated extraction of argumentative language patterns is still considered a challenging task [18]. But even if arguments were to be detected reliably, their application as explainable components in a RS framework is not trivially given.

One particular obstacle is the missing link between user preferences and argument relevance. Naturally, user opinions are multi-faceted and personal attitudes towards the different aspects of a product domain strongly contribute to their evaluation [32]. For instance, when choosing a movie to watch, the decision is presumably influenced by its genre, story, visuals, or by appearing actors etc. In addition, opin-

ions about such aspects may be conflicting. Effective persuasion is, therefore, dependent on the consideration of the target audience's perspective with respect to specific aspect categories. The identification of aspects can, in this context, be described as a form of topic modeling in which a target entity, i.e. an item of the product domain, is linked with certain attributes towards which opinions can be expressed [21]. Being able to identify arguments per se, is, under this light, only a partial solution to the provision of relevant premises. Rather, the RS has to consider which pieces of available information are likely to be deemed as important by the target user and how this information can be represented in the larger context of a decision task. An analogy can easily be drawn from real-life: When friends give recommendations to each other, they usually accompany their claim by carefully selected reasons that are targeted at their vis-à-vis. Equivalently, in an automated setting, the alignment between argument and audience is crucial as well.

As a result, we define two prerequisites for the acquisition of personally relevant arguments: First, the identification of domain aspects representing the dimensions based on which arguments can be selected. Second, the derivation of a notion of how the target user evaluates these aspects. For instance, it would not suffice to identify *story* as a salient aspect of the movie domain. Instead, it is also important to assess which kinds of stories, i.e. the concrete aspect realization, the user prefers. In order to achieve such a level of distinction, a method to detect preferential relations between users and items has to be established. For the work at hand, we assume such personalized inferences can be derived from utterances that contain indicators of polarity, i.e. positive and negative sentiment.

While we focus on developing an architecture for modeling user preferences based on review data in our work presented here, in further research in the ASSURE project, we plan to address several critical challenges entailed by the integration of argumentation principles: Although sentiment analysis has provided successful techniques in practice, they only tell what opinions have been expressed, but not why these opinions are held in the first place. Consequently, there is no guarantee that the identified statements will be argumentative. It is not uncommon for users to only state that they liked or disliked a movie without giving any reasons why. In other cases, reviews might as well be descriptive only. For example, people may describe the *story* in great detail without adding any evaluative content. To make the problem even more complex, descriptive and evaluative components might not be adjacent in text, but be separated, for instance, by punctuation marks. Coreference resolution [27], argumentative zoning [29], or reasoning about entailment [31] are only some of the techniques that can play an important role to solve this problem. Otherwise,

extracted passages may be incomprehensible due to a lack of context [5].

Moreover, argumentation mining is not only concerned with the identification of individual claims and premises being made, but also with the derivation of relationships between them and how they work together to support or undermine the overall message. The extraction of argument graphs is a powerful tool to provide users with rich explanations that shed light on an item's properties from various perspectives. Notably, the construction of argument graphs is not limited to a single review. Theoretically, one can assume a relation between the arguments being made in several reviews. Please refer to pertinent overview works to find several other current challenges of argumentation mining in general [e.g. 17].

While current argumentation mining techniques are still limited in solving the problems addressed above, some obstacles may be overcome in practice. For example, due to the large amount of available user reviews, it is not necessary to identify every single argument that could theoretically be found. It would rather be sufficient to identify a number of high quality arguments while dropping those argument candidates where the classifier is uncertain. The latter cases are often characterized by implicitness of premises and may, therefore, be hard to understand by users. Therefore, the extraction of unambiguous arguments, as indicated by, for example, discourse indicators such as *because*, might even be preferable.

3 Aspect-based Transparent Memories

As we have described in Sect. 2, the personalized extraction of polar structures is central to our purpose. Doing so requires the establishment of a user model that represents individual attitudes towards relevant aspects of the target domain. In this section, we introduce a novel method, which we call *Aspect-based Transparent Memories* (ATM), that models such multi-faceted user preferences and compares them to an item's properties in order to accurately predict numeric ratings while, at the same time, identifying candidate sentences to explain this prediction. Neural memory-based methods [cf. 10, 28] allow the externalization and structurization of possibly large amounts of knowledge. In our case, the memories are unique to a single person or item and control the process of encoding and decoding review data. Both steps, encoding (or *writing*) and decoding (or *reading*) are accompanied by mechanisms that are designed to impose transparency on the model.

The ATM architecture (Fig. 1) consists of two neural memories that can be viewed as arrays of slots for storing and thus memorizing information [10, 28]. The first memory component encodes representations of sentences

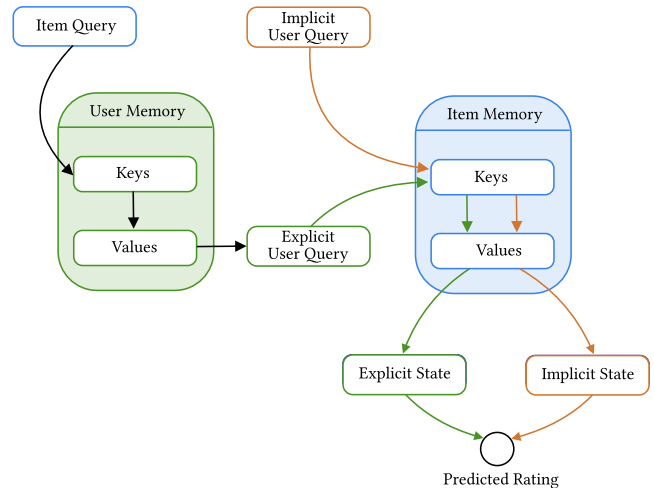


Fig. 1 Simplified schematic illustration of the proposed rating prediction pipeline including read operations for the neural memory components

composed by the target user. The second one is an equivalent variant for the target item and encompasses statements about the item by other users. Both memories are comprised of two subcomponents. First, aspect-based key vectors are used to perform the addressing operation, i.e. the selection of relevant memory locations. Keys are calculated by reconstructing sentences as a weighted combination of aspect embeddings such that the memory can be read in terms of topical overlap with the query vector. This model component is adapted from [11] and learns how to extract aspects in an unsupervised fashion while, at the same time, identifying the most salient of these learned aspects in each sentence. Conceptually, both aspect extraction as well as memory addressing can be described as a form of neural attention [2]. Value vectors, i.e. the content encoded into memory, depict the second component and contain the encoded sentence semantics. In our case, we encoded the sentences each with a bidirectional LSTM [12].

In order to predict the target rating, read operations extract elements from both user and item memory in a mixed-initiative fashion. The user memory is first queried by an item embedding, that is learned during the training process, to calculate the match between user preferences and item properties with respect to the appearing aspects. We call the result the *explicit* user state since it is derived from sentences put into writing as an active process by the user. This user state serves as a query to the aforementioned item memory. In other words, the explicit user interests are aligned with the opinions expressed by other users.

However, certain patterns found in rating behavior cannot be explained in terms of review content alone. For instance, a user may especially like *fantasy* movies, although they never mention this explicitly in any of their reviews. We assume that addressing the item memory only with the

explicit user state is insufficient. Therefore, we additionally train an *implicit* user representation that captures latent patterns similar to conventional collaborative filtering. This implicit state then serves as an additional query to the item memory. Both resulting vectors, explicit and implicit, can subsequently be combined to predict the target rating. A description of the architectural details and formalism used as well as extensions to the ATM architecture can be found in [5].

4 Explanations

Rendering recommendation models explainable has been recognized as a means to help users verify the underlying rationale by increasing transparency and accountability [e.g. 16]. Although retro-fit interpretable models have been proposed in the past [e.g. 24], we follow the line of argumentation that only model-intrinsic explanations allow faithful insights into the actual qualitative relationship between input features and recommendations [26]. Post-hoc explanations, on the other hand, cannot provide a sufficient level of certainty about their truthfulness as they usually only provide approximations of internal model states.

In order to generate human-intelligible explanations, we propose to exploit the states of ATM's diverse attention components. Accessing attention values allows us to formulate two different types of explanations: The first one deals with the problem of representing the information space from the system's perspective. By indicating which aspects the system attends to and by clarifying how these aspects relate to concrete utterances, the target user becomes empowered to assess how the system evaluates and structures the input information. The second kind of explanation is rather concerned with extracting the reasons behind one concrete recommendation. Again, the attention components can be exploited to pick statements from other users that support this claim. In the following, we describe details concerning how to arrive at these types of explanations:

Aspect Extraction. In order to provide an overview of the information space, ATM can convey details about the average distribution of aspects in the whole data set (Fig. 2) or for a specific item (Fig. 3). The first step for this is to derive which aspects are deemed important by the system in the first place. The set of attended aspects can either be fixed a-priori or learned in unison with the remaining network parameters. Fixing them can be achieved by setting the aspect representation to the word embedding of the respective aspect term or by averaging the embedding vectors of several terms that together form a higher-order aspect. For instance, in the movie domain one such combination may consist of the embeddings of *story*, *storytelling*, *script* etc.

Opposed to this, automatically identifying aspects can be achieved via extending the model cost function by an unsupervised loss that measures how well sentences can be reconstructed solely based on a combination of learned aspect embeddings [11]. For a given sentence, we can then derive the relative importance of each aspect. Consequently, averaging this importance rating over all sentences yields the overall distribution of occurring aspects in the data.

Personal Aspect Importance. Once salient aspects have been detected, ATM can utilize this information further to assess which aspects are assumed to be especially important to the target user (Fig. 2). As before, this information can be extracted by averaging the aspect weights; only this time the target sentences shall originate only from of the current user. Merely showing the distribution of personal aspect importance, however, doesn't yield sufficient transparency. Instead, we can additionally display exemplary sentences that strongly contributed to a particular aspect weight. Through this step, the user can better verify whether they agree with the assessment of their assumed aspect preferences.

Recommendation Explanation. The central explanatory component of ATM is a mechanism to communicate the reasoning behind one particular recommendation. As displayed in Fig. 1, ATM matches the user's explicit and implicit representations against statements formulated by other users. The resulting attention weights then indicate which sentences contain the largest overlap with the vector-

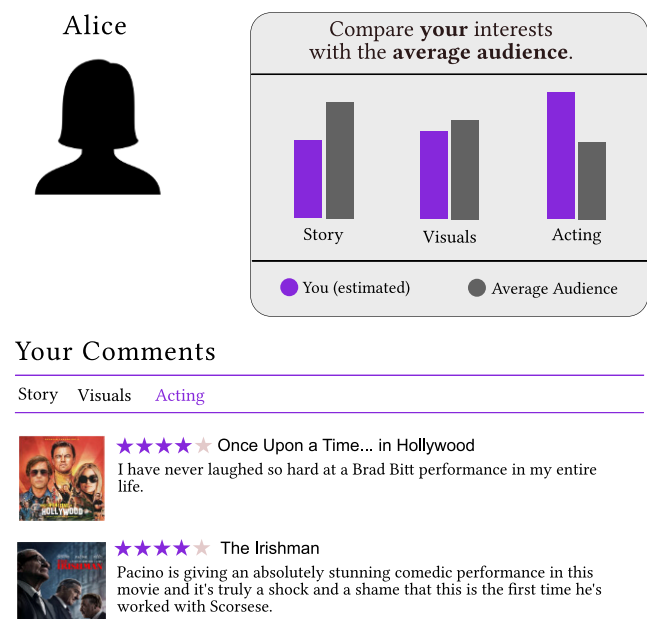


Fig. 2 Exemplary user profile that depicting (assumed) personal and average importance for three aspects as well as the target user's comments sorted by aspects

Fig. 3 Recommendation for the movie *Parasite* including the predicted rating, an overview of (assumed) aspect importance, and a personally selected comment that supports the predicted rating. Selection of comments can be personalized by toggling the respective radio button

Parasite (2019)
Directed by Bong Joon-ho

Predicted Rating: ★★★★★

This rating was predicted based on your past ratings and reviews. See personally selected comments below for reasons why we think you will like this movie.

Summary:
All unemployed, Ki-taek's family takes peculiar interest in the wealthy and glamorous Parks for their livelihood until they get entangled in an unexpected incident.

Compare what we think **you** will find interesting about *Parasite* with the **average audience**.

Aspect	You (estimated)	Average Audience
Story	★★★★★	★★★★
Visuals	★★★★	★★★★
Acting	★★★★	★★★★

Comments

Story Visuals Acting Personalized

Bob ★★★★★
Even as the plot ramps up and the tone starts to shift from dark comedy into tense thriller, Bong keeps a masterful hold on the reins.

ized user preferences. In other words, sentences with large attention weights are the best candidates to describe what properties of the target item the current user will probably like or dislike most (Fig. 3). Informally, the explanation process can be exemplified as follows: Let us assume a user has exhaustively dealt with storytelling in their past reviews. Concretely, they seem to like complex stories with a twist-ending a lot. ATM would then, based on concrete examples of these statement, derive a user representation that semantically reflects this preference. Now, if ATM were to generate recommendations for this user, it would find large overlaps between their preference representation and the embedding of sentences that also deal with twist-endings. As a result, not only will fitting movies receive larger overall scores, but individual sentences for this movie that contain concrete information about their ending would also be detected as salient. Please note that the same also applies for the implicit user representation. For a more detailed discussion of explaining recommendations with this process, please refer to our work presented in [5]. It also presents a user study aimed at evaluating the quality of the explanations generated.

Summarized, ATM can be seen as the first step towards a full-fledged argumentation-based explainable RS. In its current state, ATM is mostly concerned with detecting personally important information in review texts composed by other users. However, the extracted content is not yet presented in argumentative manner as no structural knowledge about arguments is represented in the model. This leads to several limitations that were already discussed in Sect. 2. Please note, however, that although such natural language explanations may eventually entail a causal structure, the underlying attention mechanism still only operates as a correlational statistical process. The resulting explanations, therefore, only express merely apparent causality.

This is a phenomenon that has to be further investigated in future works via, for instance, the application of causal reasoning techniques [e.g. 8].

5 Evaluation

We have conducted experiments on three real-world datasets to demonstrate the effectiveness of ATM by comparing it to state-of-the-art RS. In the following, we present the datasets used, describe our experimental procedure, and introduce the baselines selected for comparison. Then, we evaluate and discuss performance.

Datasets. The *Yelp*¹ dataset is a large-scale dataset introduced in the context of the Yelp challenge. Since aspects only relate to specific domains, we filtered out all reviews for businesses not associated with the category *Restaurants*. *Kindle* is one category of the Amazon review dataset² containing reviews of e-books purchased from the Kindle store. *Movies* is another of the Amazon categories with movie and TV reviews. All datasets contain user reviews associated with a 5-star rating.

Procedure and Settings. In our experiments, we adopted the well-known Mean Squared Error (MSE) metric to evaluate recommender performance.

Reviews were first passed through a Stanford Core NLP Tokenizer [20] to obtain tokens which were then lower-cased. Sentences were separated subject to the tokenizer result. Contractions were expanded and stopwords and punctuation were combined into a single token. We set a maxi-

¹ <https://www.yelp.com/dataset/challenge>.

² <http://jmcauley.ucsd.edu/data/amazon/>.

imum number of 30 words per sentence with a total of 150 sentences per user and item. Shorter sentences were padded accordingly. We used pretrained *fastText* embeddings [13] with dimensionality 300 for word embeddings.

We trained a variant of our proposed model with 10 aspects initialized with random values drawn from a Glorot uniform distribution [9]. The same distribution was used for randomly initializing the remaining parameters. Concerning the model-specific hyperparameters, we set $\lambda_D = 0.6$, $\lambda_s = 1.0$ and $\lambda_u = 1.0$. All of these values were selected via grid-search-like optimization.

Optimization was performed using Adam [14] and a learning rate of 0.001. We randomly split the data into training (80%), validation (10%), and test set (10%). The maximum number of epochs was set to 10. After training for one epoch with a batch size of 32, we calculated MSE on validation and test set. We report the results of the test set where the results of the validation set was lowest. All algorithms were implemented with Python using PyTorch [23].

Baselines. We compared our method against established recommendation models:

- *Matrix Factorization* (MF) [15] is one of the most popular collaborative filtering techniques.
- *Neural Attentional Rating Regression* (NARRE) [3] is a convolutional model that consists of two parallel attentive neural networks coupled by a final recommendation layer. The first network processes reviews of a target user in an attentive manner to derive a latent state. The second does the corresponding operation for the item side. Since we were mainly interested in comparing review-retrieval approaches with our aspect-based variant, we view NARRE, known to produce state-of-the-art recommendations, as a representative instance of the whole line of research.

The parameter values for NARRE were assigned subject to the evaluation in the original paper. The textual data made available to NARRE were chosen to match the total amount available to ATM.

Offline Performance. The results for the rating prediction task for ATM and the baselines are given in Table 1. As was

Table 1 MSE for all baselines as well as the proposed ATM

	Yelp	Kindle	Movies
<i>Baselines</i>			
MF	1.379	0.739	1.488
NARRE	1.102	0.519	0.922
ATM	1.085	0.454	0.847

to be expected, the two review-based approaches perform better than the conventional collaborative filtering model.

Furthermore, our proposed model outperforms NARRE on all considered datasets. One explanation for this is that its mixed-initiative approach allows ATM to more selectively distribute attention. While the integration between target user and item only happens in the later layers in NARRE, user and item side are interwoven in ATM right from the beginning. Additionally, breaking reviews down into sentences allows the model to distribute its attention on smaller semantic units. Finally, the integration of the review reconstruction pipeline strengthens the overall training signal.

6 Conclusion

In this paper, we present an overview of the ASSURE project and the role of argumentation in recommender systems. We furthermore describe in more detail one of the solutions developed in the project: ATM is an approach to memorize user opinions on relevant item aspects found in raw review texts to derive multi-faceted user and item representations. We have shown that representing knowledge about multiple aspects in combination with external memories leads to more accurate recommendations. Offline experiments indicate that ATM outperforms other review-based models by at least a slight margin. The model can also serve as a basis for generating more informative explanations. These include the arrangement of review content with respect to aspect categories as well as the provision of personally selected user comments as decision support.

However, there is still room for improvement. Since ATM currently disregards any language structure beyond discriminating sentences, this may lead to explanations being detached from their original context which, in turn, impedes intelligibility. Consequently, the incorporation of deeper linguistic preprocessing appears necessary to improve the explanation performance. We are currently extending the approach by including representations of discourse markers and more complex argumentation mining techniques to reliably detect argumentative structures. Finally, we are also investigating means of formulating multi-perspective explanations based on supporting and attacking relations as derived from argument graphs generated from review data.

Funding Open Access funding provided by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are

included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Askalidis G, Malthouse EC (2016) The value of online customer reviews. In: Proceedings of the 10th ACM Conference on Recommender Systems, ACM
2. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate (arXiv preprint arXiv:14090473)
3. Chen C, Zhang M, Liu Y, Ma S (2018) Neural attentional rating regression with review-level explanations. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web, pp 1583–1592 (International World Wide Web Conferences Steering Committee)
4. Chesnevar CI, Maguitman AG, González MP (2009) Empowering recommendation technologies through argumentation. In: Rahwan I, Simari GR (eds) *Argumentation in artificial intelligence*. Springer, Heidelberg, Berlin, New York, pp 403–422
5. Donkers T, Kleemann T, Ziegler J (2020) Explaining recommendations by means of aspect-based transparent memories. In: Proceedings of the 25th International Conference on Intelligent User Interfaces, pp 166–176
6. Gedikli F, Jannach D, Ge M (2014) How should I explain? A comparison of different explanation types for recommender systems. *Int J Hum Comput Stud* 72(4):367–382
7. Gena C, Grillo P, Lieto A, Mattutino C, Vernerio F (2019) When personalization is not an option: an in-the-wild study on persuasive news recommendation. *Information* 10(10):300
8. Ghazimatin A, Balalau O, Roy RS, Weikum G (2019) PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems (arXiv preprint arXiv:191108378)
9. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp 249–256
10. Graves A, Wayne G, Danihelka I (2014) Neural Turing machines (arXiv preprint arXiv:14105401)
11. He R, Lee WS, Ng HT, Dahlmeier D (2017) An unsupervised neural attention model for aspect extraction. In: Long Papers. Proceedings of the 55th annual meeting of the association for computational linguistics, vol 1. In, Vancouver, Canada, pp 388–397
12. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
13. Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification (arXiv preprint arXiv:160701759)
14. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization (arXiv preprint arXiv:1412.6980)
15. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. *Computer* 8:30–37
16. Kunkel J, Donkers T, Michael L, Barbu CM, Ziegler J (2019) Let me explain: impact of personal and impersonal explanations on trust in recommender systems. *CHI Conference on Human Factors in Computing Systems Proceedings*, CHI 2019, ACM, New York, NY, USA
17. Lawrence J, Reed C (2020) Argument mining: a survey. *Comput Linguist* 45(4):765–818
18. Lippi M, Torroni P (2016) Argumentation mining: state of the art and emerging trends. *ACM Trans Internet Technol* 16(2):10
19. Loepp B, Donkers T, Kleemann T, Ziegler J (2019) Interactive recommending with Tag-enhanced Matrix Factorization (TagMF). *Int J Hum Comput Stud* 121:21–41
20. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp 55–60
21. McAuley J, Leskovec J, Jurafsky D (2012) Learning attitudes and attributes from multi-aspect reviews. In: 2012 IEEE 12th International Conference on Data Mining, IEEE, pp 1020–1025
22. Naveed S, Donkers T, Ziegler J (2018) Argumentation-based explanations in recommender systems: conceptual framework and empirical results. In: Adjunct publication of the 26th Conference on User Modeling, Adaptation and Personalization, ACM UMAP '18, Singapore, pp 293–298
23. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in pytorch
24. Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 1135–1144
25. Ricci F, Rokach L, Shapira B (2015) Recommender systems: introduction and challenges. In: *Recommender systems handbook*. Springer, Heidelberg, Berlin, New York, pp 1–34
26. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1(5):206
27. Soon WM, Ng HT, Lim DCY (2001) A machine learning approach to coreference resolution of noun phrases. *Comput Linguist* 27(4):521–544
28. Sukhbaatar S, Weston J, Fergus R (2015) End-to-end memory networks. In: *Advances in neural information processing systems*, pp 2440–2448
29. Teufel S, Siddharthan A, Batchelor C (2009) Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, vol 3. Association for Computational Linguistics, Singapore, pp 1493–1502
30. Tintarev N, Masthoff J (2011) Designing and evaluating explanations for recommender systems. In: *Recommender systems handbook*. Springer, Heidelberg, Berlin, New York, pp 479–510
31. Zanzotto FM, Pennacchiotti M, Moschitti A (2009) A machine learning approach to textual entailment recognition. *Nat Lang Eng* 15(4):551–582
32. Zhang Y, Lai G, Zhang M, Zhang Y, Liu Y, Ma S (2014) Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In: Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval, ACM, pp 83–92

The following article is reused from:

Donkers, T., & Ziegler, J. (2021). The Dual Echo Chamber: Modeling Social Media Polarization for Interventional Recommending. In *Proceedings of the 15th ACM Conference on Recommender Systems* (pp. 12-22). Association for Computing Machinery.
<https://doi.org/10.1145/3460231.3474261>



The Dual Echo Chamber: Modeling Social Media Polarization for Interventional Recommending

Tim Donkers
tim.donkers@uni-due.de

University of Duisburg-Essen
Duisburg, North Rhine-Westphalia, Germany

Jürgen Ziegler
juergen.ziegler@uni-due.de

University of Duisburg-Essen
Duisburg, North Rhine-Westphalia, Germany

ABSTRACT

Echo chambers are social phenomena that amplify agreement and suppress opposing views in social media which may lead to fragmentation and polarization of the user population. In prior research, echo chambers have mainly been modeled as a result of social information diffusion. While most scientific work has framed echo chambers as a result of epistemic imbalances between polarized communities, we argue that members of echo chambers often actively discredit outside sources to maintain coherent world views. We therefore argue that two different types of echo chambers occur in social media contexts: Epistemic echo chambers create information gaps mainly through their structure whereas ideological echo chambers systematically exclude counter-attitudinal information. Diversifying recommendations by simply widening the scope of topics and viewpoints covered to counteract the echo chamber effect may be ineffective in such contexts. To investigate the characteristics of this dual echo chamber view and to assess the depolarizing effects of diversified recommendations, we apply an agent-based modeling approach. We rely on knowledge graph embedding techniques not only to generate recommendations, but also to show how to utilize logical graph queries in embedding spaces to diversify recommendations aimed at challenging polarization in online discussions. The results of our evaluation indicate that counteracting the two different types of echo chambers requires fundamentally different diversification strategies.

CCS CONCEPTS

• **Computing methodologies** → Reasoning about belief and knowledge; Logical and relational learning.

KEYWORDS

recommender systems, knowledge graphs, agent-based modeling, machine learning

ACM Reference Format:

Tim Donkers and Jürgen Ziegler. 2021. The Dual Echo Chamber: Modeling Social Media Polarization for Interventional Recommending. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21)*, September 27-October

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

RecSys '21, September 27-October 1, 2021, Amsterdam, Netherlands

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8458-2/21/09... \$15.00

<https://doi.org/10.1145/3460231.3474261>

1, 2021, Amsterdam, Netherlands. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3460231.3474261>

1 INTRODUCTION

Digital technologies have enabled people to easily connect and organize based on shared interests, even over long distances. Nowadays, however, many people do not merely engage in social exchange on the internet, but use social networks as a source for information search as well [1, 67]. Several researchers have voiced concern that the tendency to connect to like-minded individuals restricts the size of the accessible argument pool, thus exposing people exclusively to information and opinions that reinforce their existing beliefs while isolating them from potentially dissenting views [43, 48]. According to Sunstein, the modal outcome of this process is the circulatory amplification and subsequent radicalization of standpoints as well as the segregation of the public space into disconnected spheres shaped by a lack of consensus [61]. This phenomenon has come to be known as the *echo chamber effect* [5, 6, 38, 61, 62].

It remains disputable if and under which circumstances digital technologies may have such drastic negative effects on a society. Barberá, for instance, observes that cross-cutting interaction is still a frequent phenomenon in the digital domain and that exposure to diverse news appears to be higher than for traditional media [5]. Some authors derive from this that the echo chamber phenomenon would not exist at all [11, 12, 21]. However, these works focus primarily on measuring connectivity and exposure on social media, thus framing the echo chamber effect as an epistemic imbalance across communities. What they miss to account for is the fact that exposure to information is not to be equated with its integration into one's personal belief system.

Psychological research suggests that human decision-making processes are affected by cognitive biases [22]. For instance, people tend to engage in selective exposure seeking out information that confirms their own world view in order to maintain a cognitive equilibrium [23]. Jamieson and Cappella argue that the echo chamber phenomenon encompasses mechanisms that go even further than that [38]. According to them, echo chambers work by systematically alienating members of a community from outside epistemic sources. Such dysfunctional ideological patterns not only prevent people from engaging in informative search beyond their intellectual community, but, even more importantly, set the motives to actively discredit outside voices [63]. Where an epistemic imbalance would imply the mere omission of contrary views, ideological preemptive distrust may give rise to collective foreclosure. Therefore, we argue that echo chambers are often social structures systematically excluding sources of information not necessarily by omission,

but through deliberate action. Put together, in this work we distinguish two types of echo chambers with systematic differences: *Epistemic* echo chambers (also referred to as epistemic bubbles [68]) mainly yield informational gaps through an impairment of their structure whereas *ideological* echo chambers systematically exclude counter-attitudinal information [51].

The impact of human behavior on social fragmentation is, however, only one side of the coin. The structural inhibition of information diffusion may also be accounted for by algorithmic biases of the social networking site itself. Since the overabundance of information would otherwise be overwhelming to people, social media platforms are forced to outsource the filtering process to recommender systems. The underlying algorithms are designed to personalize the stream of information in order to pick the most relevant content. However, social networks require an adequate amount of diversity to allow for informational sovereignty of their users. Unfortunately, imposing personalized filters on information may exacerbate fragmentation by creating degenerate feedback loops where broadness of information gets increasingly pruned over time [39].

Systemic influence of personalized information streams is usually discussed in the context of *filter bubbles* [55] defined as an individual outcome of different algorithmic processes on the system side and cognitive processes such as information search and perception on the user side. As a result, the information made available is strongly tailored towards pre-existing attitudes and beliefs. In scientific literature, however, both terms have often been used interchangeably although echo chambers describe the overall phenomenon while filter bubbles rather refer to self-reinforcing patterns of information reduction caused by algorithmic influence. As a result, respective empirical evidence concerning the role of technology is notably sparse [5].

We take this as motivation to systematically investigate the reciprocal relationship between user and system in light of the echo chamber phenomenon. For this purpose, we have created a simulation paradigm that involves both user and system as two agents influencing each other. We are specifically interested in capturing the impact of varying algorithmic influence in both the epistemic and the ideological scenarios. We propose an *agent-based modeling* [44] procedure with which we can parametrize two types of user behavior: In the epistemic condition, users will select content with respect to preferential dispositions, i.e. close sources are preferred to more distant ones. However, the decision process is not influenced by the identification with beliefs typical for the respective echo chamber. In the ideological condition, on the other hand, influence of outsiders is diminished mimicking discrediting behavior. We generally assume that the closer a user is to the ideological center of a community, the stronger they will value the ideological proximity towards potential sharers of information.

In order to adequately represent social network structures, we utilize graph-based recommendation techniques, namely *knowledge graph embeddings* [14]. Knowledge graph embeddings are an active field of research with notable contributions to link prediction [9], question answering [8], and recommender systems [32, 53, 54, 71] in recent years. Our goal is to project entities into a latent embedding space based on information extracted from heterogeneous edges in a social network graph. The resulting embeddings can subsequently

be used to generate personalized recommendations. In addition, knowledge graph embeddings offer elevated levels of control over the embedding space via the application of logical graph queries that meaningfully project entities into different areas [34, 57]. We exploit this property not only to formulate diversification principles for recommendations, but also show that we can use rather simple queries to parametrize human cognitive filters.

We then bring both technological and human filters together in a single framework: Based on a real-world empirical dataset scraped from Twitter, we monitor the evolution of an online discussion over time and intervene at specific points to simulate user interactions with a new set of synthetic recommendations generated by technological filters. Although this setup does explicitly not resemble real-world interactions, we nonetheless argue that valuable insights can be gained from such a procedure. As an analogy, consider a laboratory study where active users are invited to engage with the available information from a new perspective; only this time under controlled circumstances and inspection.

Based on well-established community detection techniques [17], we verify that densely packed accumulations of users can, indeed, be identified. In order to challenge fragmentation, we find that epistemic and ideological chambers require fundamentally different interventions. While epistemic chambers can easily be dissolved via the establishment of new (random) connections, more targeted operations are required in the ideological case. We argue that ideological chambers are such a robust phenomenon that their homogeneity is best tackled by highlighting critical intra-community discourse. An intervention can be considered effective if content of community members is proposed who engage with contradictory ideological positions, critically reflect their own views and especially share serendipitous positions [18, 64]. Therefore, we develop methods that project users into border areas between communities where most of the cross-ideological exchange takes place.

Summarized, our contributions are as follows: We formulate an agent-based modeling framework that we enrich with knowledge graph embeddings derived from social media networks. We show that logical operators applied to latent embedding spaces are powerful tools capable of modeling complex technological as well as cognitive phenomena. Concretely, we use these operators to calculate recommendations but also diversify them meaningfully according to the restrictions imposed by the echo chamber phenomenon. We additionally show, by devising two different cognitive acceptance processes subject to epistemic and ideological scenarios, that distances in the embedding space can be exploited to model cognitive filters. While our evaluation on real-world data indicates that both scenarios follow comparable dynamics when modeling echo chambers as a process of local information distribution, notably different patterns can be observed when recommendations are diversified. Our results, therefore, stress that evaluating the effectiveness of depolarization techniques has to be undertaken with respect to carefully designed human decision procedures.

The remainder of this paper is structured as follows: In Section 2, we give an overview of the existing research on the echo chamber effect and highlight technological approaches to depolarization. In Section 3, the algorithmic foundation of our framework is described. Concretely, we derive the basic principle of conducting conjunctive graph queries in an embedding space. Section 4 focuses on

the implementation of the agent-based modeling procedure with reference to knowledge graph embeddings. Section 5 provides the results of our evaluation which we finally discuss in Section 6.

2 RELATED WORK

It has long been known that homophilic interaction causes people to adopt more extreme positions [49]. Although online spaces create the opportunity for enclave deliberation, Sunstein argues that, in practice, the outcome of such processes is group polarization as they represent a “*breeding ground for extremism*” [61, 62]. Accordingly, Del Vicario et al. observe that information related to scientific news and conspiracy theories tends to spread in homogeneous and polarized communities on Facebook [20]. However, it is still mostly unclear whether homophilic interaction patterns are related to one’s ability to filter out information challenging pre-existing attitudes. In this regard, Garrett demonstrates that, while people tend to increase exposure to political stories consistent with their views, they do not systematically avoid opinion challenges [26]. Rather, as Brundidge puts it, people become inadvertently exposed to (political) differences through a weakening of social boundaries in online environments [10]. Both Bakshy et al. [4] as well as Barberá et al. [6] independently report that around 30% of the political news stories on Facebook and Twitter are cross-cutting.

Nonetheless, most attempts to simulate the emergence of echo chambers act on the assumption that social fragmentation is a result of limited exposure due to local information distribution. The underlying mechanisms of polarization have been studied with respect to several mathematical models, such as opinion dynamics [36], social influence [24] or cultural dissemination [2]. Dandekar et al., for instance, generalize DeGroot’s model of opinion formation [19] to account for subjective biased assimilation and show that biased individuals suffer from increased polarization tendencies [15]. Del Vicario et al. model confirmation bias and polarization subject to a bounded confidence model [20]. Geschke et al. apply agent-based modeling to capture the dynamics that emerge during user-to-user interaction on social platforms [28]. Similarly, Sasahara et al. rely on advances in opinion dynamics to model mutual influence and unfriending behavior to verify the development of separated communities [58].

The decision-making process of whether to integrate a proposed message is usually modeled as a function of proximity in an attitude space [7, 15, 28, 58]. In order to challenge fragmentation, the hope is to suggest more diverse content that should lead to an alignment of opinions over time. Although the modeling of cognitive biases does receive attention [15, 28], the specific case of ideological segregation is usually disregarded. One notable exception is the work by Wang et al. who find that ideological reference points are a key determinant of network polarization [65]. Additionally, other studies indicate that in-out-group dynamics likely cause complex interactions between people subject to their community membership and, hence, ideological position [49, 59].

The works described above share the assumption that network polarization can be modeled as a largely social process. In the few cases where technological influence is taken into account, recommendation generation is only implemented in a rudimentary way [e.g. 15, 28]. In his influential book “*The Filter Bubble*” [55], Eli

Pariser describes the impact of real-time, automated, and personalized algorithms on the process of information diffusion. Unfortunately, isolating the influencing factors of technology turns out to be difficult as researchers rarely have access to proprietary search and ranking algorithms. Unsurprisingly, the most extensive study in this regard was conducted by Facebook: Bakshy et al. show that Facebook’s algorithm significantly reduces user exposure to cross-cutting content [4]. However, only small effect sizes can be reported.

As one of the few examples that account for algorithmic influence, Badami et al. introduce a diversification approach to rating-based recommender systems designed to combat over-specialization in polarized environments [3]. However, the situation for e-commerce recommender systems is quite different in nature to our problem of interest as rating data is usually not available on social media.

Due to the limited consideration of technological influence, system-side techniques aimed at challenging polarity have rarely been proposed as well. Conceptually closest to our work is a paper by Garimella et al. who propose a mechanism to reduce polarization by connecting opposing views [25]. Grevet et al. find that selective exposure can be mitigated by exposing users to weak ties [31]. Liao et al. demonstrate that appending source position indicators increases the likelihood of users selecting moderate content. Graells-Garrido et al. state that content diversity is of utmost importance to challenge polarization [30]. As a solution to increase diversity, they propose to find intermediary topics by constructing a topic graph.

In terms of the underlying recommendation technology, our method is related to advances in graph-based recommendations. In recent years, research has provided significant progress in using machine learning to reason with relational data, especially within the context of knowledge graph embeddings [14]. While early works have primarily focused on the representation problem of graphs in latent information spaces [9, 66], over time new applications such as recommender systems have been explored [53, 54, 71]. For an overview of knowledge graph based recommender systems, please refer to [32]. Respective methods are mainly build to increase prediction accuracy, whereas our aim is rather to explore mechanisms to increase intra-network connectivity. We intend to achieve this goal by exploring the embedding space to suggest new connections between users that go beyond local neighborhoods. Hence, the strongest inspiration to our work is the concept of logical reasoning in embedding spaces [16, 34, 57]. Particularly influential is a paper by Hamilton et al. who propose an approach to make predictions about conjunctive graph queries [34].

3 KNOWLEDGE GRAPH EMBEDDINGS

In this section, we lay out the foundation of our agent-based modeling procedure by formulating an embedding-based framework to efficiently make predictions on incomplete graphs. The central idea is that we embed graph nodes into a low-dimensional space and represent logical operators as geometric operations in this embedding space. After training, we use this model to predict which nodes are likely to satisfy our proposed queries.

To give a concrete example, given an interaction graph of a social media platform, we may pose the conjunctive query “*return all messages composed by user a that user b will likely engage with.*”

Queries can also take on more complex structures such as “*return all messages composed by user a that user b will likely engage with and, at the same time, will be shared in community c.*”

We consider knowledge graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of nodes $v \in \mathcal{V}$ and directed edges $e \in \mathcal{E}$. We denote edges as binary predicates $e = \tau(u, v)$ with $\tau \in \mathcal{R}$ and $u, v \in \mathcal{V}$. Each node is associated with a type $\gamma \in \Gamma$ such that $\tau : \gamma_1 \times \gamma_2 \rightarrow \{0, 1\}$ describes the edge relation.

One important property of knowledge-graph embeddings is the possibility to not only discover relations actually contained in the graph, but also to predict unobserved but likely connections. This depicts the primary functionality to generate recommendations. Formally, we assume that every query $q \in Q(G)$ has some unobserved denotation set $\llbracket q \rrbracket$ that we are trying to predict. Please note that it is likely that $\llbracket q \rrbracket$ is not fully contained in our training data.

In order to represent logical query operations in latent space, we follow [34] and map the conjunctive input query q to an embedding. This process is realized in three steps: We (i) derive a set of anchor nodes based on the query’s dependency graph, (ii) embed them into latent space, and (iii) apply geometric operators until we reach the query target. Thereby, node embeddings $\mathbf{v} \in \mathbb{R}^d$ are learned as a structural node property. We employ two different geometric operators \mathcal{P} and \mathcal{I} that are translated into differentiable form as well and are optimized along the embeddings for graph nodes.

The geometric translation operator \mathcal{P} outputs a new query embedding $\mathbf{q}' = \mathcal{P}(\mathbf{q}, \tau)$, given a query embedding \mathbf{q} and an edge type τ . The corresponding denotation set is defined as $\llbracket q' \rrbracket = \bigcup_{v \in \llbracket q \rrbracket} N(v, \tau)$ where $N(v, \tau)$ is the set of nodes connected to v by edges of type τ . Hence, \mathcal{P} takes an embedding corresponding to $\llbracket q \rrbracket$ and produces a new one that represents the union of all neighbor nodes in $\llbracket q \rrbracket$ that exhibit τ . This translation operation is implemented following a long line of work on encoding edge and path relations in knowledge graphs [9]:

$$\mathcal{P}(\mathbf{q}, \tau) = \mathbf{q} + \mathbf{r}_\tau, \quad (1)$$

where $\mathbf{r}_\tau \in \mathbb{R}^d$ is a trainable relation embedding for edge type τ .

Suppose we are given a set of embeddings $\{\mathbf{q}_1, \dots, \mathbf{q}_n\}$, all of which correspond to queries with the same output node type γ . The geometric intersection operator \mathcal{I} takes this set and produces a new one \mathbf{q}' with denotation $\llbracket q' \rrbracket = \bigcap_{i=1, \dots, n} \llbracket q \rrbracket_i$. That is, it performs set intersection in embedding space. We implement \mathcal{I} following [57] as:

$$\mathcal{I}(\{\mathbf{q}_1, \dots, \mathbf{q}_n\}) = \sum_{i \in \{1, \dots, n\}} a_i \mathbf{q}_i, \quad a_i = \frac{\exp(\text{NN}_k(\mathbf{q}_i))}{\sum_j \exp(\text{NN}_k(\mathbf{q}_j))} \quad (2)$$

where $a_i \in [0, 1]$ is the attention for \mathbf{q}_i and NN_k is a k -layer feed-forward neural network.

Finally, we estimate whether $v \in \llbracket q \rrbracket$ by the distance between query and node embeddings [69]:

$$d(\mathbf{q}, \mathbf{v}) = \frac{1}{2} \frac{\text{var}(\mathbf{q} - \mathbf{v})}{\text{var}(\mathbf{q}) + \text{var}(\mathbf{v})} \quad (3)$$

The embedding \mathbf{q} , thus, implicitly encompasses the denotation $\llbracket q \rrbracket$ such that $d(\mathbf{q}, \mathbf{v}) = 0, \forall v \in \llbracket q \rrbracket$ and $d(\mathbf{q}, \mathbf{v}) = 1, \forall v \notin \llbracket q \rrbracket$ is approximated. During inference, we can perform nearest neighbor search to find nodes that satisfy the query.

For a given set of queries and their respective answers, both types of operators, \mathcal{P} and \mathcal{I} , as well as node embeddings \mathbf{v} can be trained using stochastic gradient descent or a variant on a negative sampling loss [47]:

$$L(q) = -\log \sigma(\gamma - d(\mathbf{q}, \mathbf{v})) - \sum_{i=1}^k \log \sigma(d(\mathbf{q}, \mathbf{v}'_i) - \gamma), \quad (4)$$

where γ depicts a fixed scalar margin, σ is the sigmoid function, and k is the negative sample size.

4 AGENT-BASED MODELING

In this section, we describe our agent-based modeling procedure. We begin by splitting a Twitter data set into a number of m temporal bins. At every time step $k \leq m$, our training data consists of the union of edges until k : $\mathcal{E}_{0:k} = \bigcup_{i=1 \dots k} \mathcal{E}_i$. We use this data to train a knowledge graph embedding model based on which we, subsequently, formulate an intervention. This intervention simulates the interaction of users with recommendations generated by our method. Concretely, users retweet recommendations subject to a probabilistic decision procedure (defined below). The resulting simulated edges are added to the original training data for the next time period.

This means that the training procedure gets increasingly influenced by agent-based modeling and should deviate from the results gained when training the model without any intervention. Please refer to Algorithm 1 for a complete overview of the process. In the following, we clarify some preliminaries and afterwards lay out our human and technological filters.

4.1 Preliminaries

Let $\mathcal{V}_u \subset \mathcal{V}$ be the set of users and $\mathcal{V}_t \subset \mathcal{V}$ be the set of Tweets. Based on the edges until the current temporal bin $\mathcal{E}_{0:k}$, we learn to embed relations between these entities as deductive or inductive knowledge search. Some links can easily be extracted from the graph data itself. A user $u \in \mathcal{V}_u$ is connected to Tweet $t \in \mathcal{V}_t$ if they have retweeted it or composed it themselves. Intuitively, there is a relationship between users u and v when u retweets a tweet posted by v . Formally, the retweet relation is defined as $\tau_r(u, t)$. In case of the compose relation, we write $\tau_t(t, u)$. Note that for τ_t we use u as the tail component. With this, we can easily model connections between users as a transitive relation over tweets ($\tau_u(u, v) = [\tau_r(u, t), \tau_t(t, v)]$) that is realized via $\mathcal{P}(\mathcal{P}(\mathbf{u}, \tau_r), \tau_t)$ where $\mathbf{u} \in \mathbb{R}^d$ is the embedding for user u .

We also integrate a mapping between communities and users. Let $\mathcal{V}_c \subset \mathcal{V}$ be the set of identified communities. We define $\tau_c(c, u)$ as community edges meaning that there exists a relation from $c \in \mathcal{V}_c$ to u if u is part of community c . The community is chosen as the head component because this allows us to query the user space from communities. In terms of the projection operator \mathcal{P} , the query is translated to an area in latent space where nearest neighbor search would yield the most representative users of a community. Hence, we define the result of $\mathcal{P}(c, \tau_c)$ as the prototype-user of community c . Please note that communities are not explicitly represented in the graph, but are instead added inductively via the Louvain algorithm [17]. Summarized, $\{\tau_t, \tau_r, \tau_u, \tau_c\} \subset \mathcal{R}$ depicts the base set of edges

Algorithm 1: The agent-based modeling algorithm.

```

Data:  $\mathcal{V}$ ,  $\mathcal{E}$ , numEpisodes, techFilter, humanFilter
1  $\mathcal{E}_s = \emptyset$ ;
2 for  $i \leftarrow 1$  to numEpisodes do
3    $\mathcal{M} \leftarrow KGEModel$ ;
4    $\mathcal{G} = (\mathcal{V}_{0:i} \subseteq \mathcal{V}, \mathcal{E}_{0:i} \subseteq \mathcal{E})$ ;
5    $\mathcal{G}_u = (\mathcal{V}_u, \mathcal{E}_u)$  with  $\mathcal{V}_u = \{u | u \in \mathcal{V}_{0:i} \wedge y_u = user\}$  and  $\mathcal{E}_u = \{\tau_u(u, v) | u, v \in \mathcal{V}_u\} \subseteq \mathcal{E}_{0:i}$ ;
6    $(\mathcal{V}_c, \mathcal{E}_c) \leftarrow CalculateCommunities(\mathcal{G}_u)$ ;
7    $\mathcal{G} \leftarrow (\mathcal{V}_{0:i} \cup \mathcal{V}_c, \mathcal{E}_{0:i} \cup \mathcal{E}_c \cup \mathcal{E}_s)$ ;
8    $\mathcal{M} \leftarrow Train(\mathcal{M}, \mathcal{G})$ ;
9   for  $u \in \mathcal{V}_u$  do
10     $\mathcal{V}_{u^*} \subseteq \mathcal{V}_u \leftarrow RecommendUsers(u, \mathcal{M}, techFilter)$ ;
11     $S = \{(u^*, t^*, u_{t^*}) | u^* \in \mathcal{V}_{u^*} \wedge u_{t^*} \in \mathcal{V}_u \wedge t^* \in N(u^*, \tau_r) \cup N(u^*, \tau_t) \cap N(u_{t^*}, \tau_t)\} \leftarrow RecommendTweets(u, \mathcal{V}_{u^*}, \mathcal{M})^1$ ;
12    for  $(u^*, t^*, u_{t^*}) \in S$  do
13      if  $AcceptRecommendation(u, u^*, t^*, humanFilter)$  then
14        if  $\tau_u(u, u_{t^*}) \notin \mathcal{E}_{0:i} \cup \mathcal{E}_s$  then
15           $\mathcal{E}_s \leftarrow \mathcal{E}_s \cup \{\tau_u(u, u_{t^*})\}^2$ 
16        end
17         $\mathcal{E}_s \leftarrow \mathcal{E}_s \cup \{\tau_r(u, t^*)\}$ 
18      end
19    end
20 end
21 end

```

used and extended throughout our study to generate technological as well as individual human filters.

4.2 Human Filters

Each user is assigned an individual decision procedure that encompasses the aforementioned cognitive filters (see Section 1). In other words, subject to the respective scenario, users act as if they were part of either an epistemic or an ideological echo chamber. A recommendation is integrated with a certain probability based on the distance in latent space as an operationalization of such filters. With integration, we mean the acceptance of content in the sense of further distributing it via retweeting. Subject to the operationalization of individual filters, it becomes more or less likely for a user to integrate a message. This concept is known as *latitude of acceptance* in social judgment theory [60] or as *bounded confidence* in opinion dynamics literature [36].

Epistemic Echo Chambers. Whenever a new bit of information comes to the attention of an individual, we apply the bounded confidence model to decide whether a link will be created. We model the integration of information as a probabilistic event: Given a user u and a tweet t , the integration probability is a function of their distance:

$$P_e(\tau_r(u, t)) = \frac{\lambda^\delta}{d(\mathcal{P}(\mathbf{u}, \tau_r), \mathbf{t})^\delta + \lambda^\delta}, \quad (5)$$

where $\lambda \in [0, 1]$ is the latitude of acceptance, and $\delta \in \mathbb{N}$ is a sharpness parameter that determines how steep the integration probability drops from one to zero around the latitude of acceptance. Instead of directly calculating the distance between entity

¹For tweets $u^* = u_{t^*}$ and for retweets (usually) $u^* \neq u_{t^*}$. That is, in case of tweets, the candidate user is equal to the original author of a message.

²While u^* is referenced to model recommendation acceptance, simulated edges are established with u_{t^*} . We consider this a valid modeling decision as u^* adopts a mediating role between the active user u , the recommended posting t^* , and, thereby, the original author u_{t^*} .

embeddings, we utilize the projections learned from our knowledge graph embeddings.

The functional form of bounded confidence presented here goes back to [37]. Note that the probability decreases with increasing $d(\mathcal{P}(\mathbf{u}, \tau_r), \mathbf{t})$. That is, information fitting an individual's pre-existing attitudes is more likely to be integrated which resembles homophilic behavior. We follow [28] and assume constant latitude of acceptance and sharpness.

Ideological Echo Chambers. In the ideological scenario, we extend the concept of bounded confidence to account for ideological predispositions. We assume that the closer a user is to the ideological center of their chamber, the higher they value the ideological alignment with the source. Formally, we introduce an ideological scaling parameter $\epsilon \in [0, 1]$:

$$\epsilon = \frac{2d(\mathbf{u}, \mathcal{P}(c_u, \tau_c))^\kappa}{d(\mathbf{u}, \mathcal{P}(c_u, \tau_c))^\kappa + d(\mathbf{u}, \mathcal{P}(c_v, \tau_c))^\kappa} \quad (6)$$

where $\kappa \in \mathbb{N}$ is an ideological sharpness hyperparameter, c_u and c_v are the communities of target user u and user v respectively. The original latitude of acceptance is then updated as:

$$P_i(\tau_r(u, t)) = \frac{\epsilon \lambda^\delta}{d(\mathcal{P}(\mathbf{u}, \tau_r), \mathbf{t})^\delta + \epsilon \lambda^\delta}, \quad (7)$$

Equation 7 indicates how users weigh their own beliefs against those assumed by other people. Please note that $\lim_{\epsilon \rightarrow 1} P_i(\tau_r(u, t)) = P_e(\tau_r(u, t))$. That is, ideologically unbiased users integrate messages with respect to preferential dispositions only. This also applies to the case when users assess content from their own community.

4.3 Technological Filters

Technological filters cause a reduction of social connectivity by personalizing information streams. For instance, on Twitter recommendations are generated by referencing a candidate user set identified through structural links in the network graph. Concretely, recommendations are drawn from the set of postings with which

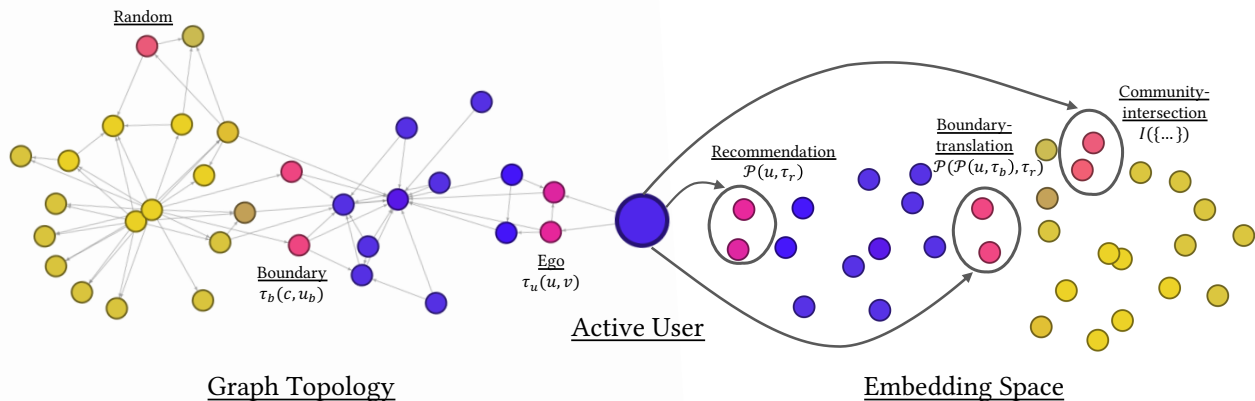


Figure 1: Schematic overview of the proposed interventions.

the candidate users have interacted in the past. We define three baselines, namely *Ego-network*, *Random*, and *Recommendation*, designed to either mimic behavior of proprietary systems, i.e. Twitter, or act as baselines to verify the effectiveness of our proposed depolarization techniques *Boundary*, *Boundary-translation*, and *Community-intersection* (see Figure 1).

While there exists a second technological filter, the ranking of items according to some objective, it is not the focus of the work at hand. Therefore, once the candidate set of items has been identified, we simply rank it in conventional recommender systems fashion via $\mathcal{P}(\mathbf{u}, \tau_r)$ and select the top- n nearest neighbors to propose to the target user.

Baselines. In social networks like Twitter, individuals are connected through one-directional follow relationships. For any given user, the platform’s underlying algorithms find the content to recommend by referencing these directly connected users and the postings they produce or distribute. We aim to approximate such standard social network behavior in a condition that we call *Ego-network*. In a second baseline, we employ a *Random* selection strategy to verify that our proposed filters do indeed exploit the structural links between entities and not merely benefit from increased diversity. Our final baseline condition, *Recommendation*, is comparable to the *Ego-network* variant in that it considers local neighborhoods to identify candidate users. Instead of relying on graph edges, however, we inspect the neighborhood in latent space. For this, we apply $\mathcal{P}(\mathbf{u}, \tau_u)$ as a user recommendation process.

Boundary. Due to the ideological structure of many echo chambers, it is unlikely that outside voices will be integrated straightforwardly. This is assumed to be especially true for individuals who identify strongly with an incompatible belief system. Therefore, we argue, any mitigating attempt should arise from an echo chamber itself. Some authors claim that changes in attitude are most likely to be caused by peers that express surprising stances [18, 64]. Our hypothesis is that statements that challenge the inherent conception of a belief system are identifiable in the boundary regions between communities where external content is introduced to insider discussions. We can easily determine the set of boundary users who have links leading inside as well as outside the community. Given

a target user u and their respective community c_u , we define candidate users as the boundary entities v_b that satisfy an edge relation $\tau_b(c_u, v_b)$.

Boundary-translation. Connecting users with boundary representatives of their community will likely increase the degree of cross-cutting interaction by confronting users with outside perspectives more frequently. However, doing so will also open the door for causing backfire-effects [52]. The mere perception of counter-attitudinal information may initiate cognitive processes aimed at preventing negative outcomes, for instance by engaging in cognitive dissonance coping. We try to reduce the probability of such effects to occur. One way to achieve this is to select only those boundary users expected to have a positive impact. Concretely, we combine the idea of user *Recommendation* and *Boundary* user selection. Given a user u and their community c_u , we establish edges between u and boundary users v_b in c_u . From this, we learn $\mathcal{P}(\mathbf{u}, \tau_b)$ and extend the original recommendation generation $\mathcal{P}(\mathbf{u}, \tau_r)$ such that it follows the path over boundary users. The resulting query is depicted as $\mathcal{P}(\mathcal{P}(\mathbf{u}, \tau_b), \tau_r)$. The argument behind this is that this translation operation identifies boundary users who are best suited to accurately predict a user’s interaction behavior.

Community-intersection. Finally, we are also interested in the effectiveness of connecting users across different communities. The underlying idea is that we want to specifically bridge polar opposites in order to depolarize the whole discussion space. To achieve this, we apply an intersection operation that returns those users from a community c that the target user would most likely connect with. Please note that we may not observe any valid edges in our training data that satisfy a particular query. However, the strength of our applied approach is that it is able to approximate likely edges not found in reality. The intersection operation is defined as $I(\{\mathcal{P}(\mathbf{c}, \tau_c), \mathcal{P}(\mathcal{P}(\mathbf{u}, \tau_r), \tau_t)\})$. Given the community c_u of user u , we find the most distant community based on $d(\mathcal{P}(c_u, \tau_c), \mathcal{P}(c_v, \tau_c))$ and calculate the intersection between the target user and the identified community.

5 EVALUATION

Based on agent-based modeling described in the previous section, we build an evaluation framework to investigate the effectiveness of our proposed interventions. We follow the process layed out in Algorithm 1 to train our recommendation model as well as to model user activities. After each intervention step, we analyze the current network state subject to dependent variables measuring fragmentation and polarization. As independent variables, we define *human* as well as *technological filters* resulting in a 2×6 study design.

As the data basis we choose the Twitter data set proposed in [29]. Included are tweets that use at least one of the following hashtags: *#blacklivesmatter*, *#alllivesmatter*, or *#bluelivesmatter*. We expect identifiable communities to form around these hashtags. Note that we randomly exclude tweets for *#blacklivesmatter* to match the number of tweets for *#alllivesmatter* and *#bluelivesmatter* combined. Furthermore, we only consider messages composed in 2020 to limit the time frame. In order to reduce data sparsity, we include user entities with, at least, 5 retweets or 3 composed tweets. Concerning tweets, we set the minimum number of retweets to 5 as well. The final dataset is comprised of 12, 548 Tweets, 21, 061 users, and 130, 780 retweet edges. Based on pre-tests, we have verified that the data set exhibits a sufficiently polarizing structure.

We split our training data into 10 bins. During each iteration and for a given target user, we initially select 100 individuals as the *user candidates* $\mathcal{V}_{u^*} \subset \mathcal{V}_u$ drawn subject to technological filters (see Section 4.3). In a next step, *item candidates* are determined by $\mathcal{V}_{t^*} = \bigcup_{u \in \mathcal{V}_{u^*}} N(u, \tau_r) \cup N(u, \tau_t)$. That is, we define as the candidate item set all items either composed or retweeted by candidate users. From \mathcal{V}_{t^*} the 20 items with the highest rank are selected as recommendations proposed to the active user. The integration of items is modeled with respect to human filters (see Section 4.2). Regarding the hyperparameters for the bounded confidence model, we set $\lambda = 0.3$ and $\delta = 5$. We choose $\kappa = 5$ in the ideological condition.

We initialize the knowledge graph embedding model with input dimensionality equal to the number of entities at time k , hidden dimensionality $d = 300$, and $\gamma = 20$. We use a single-layer neural network to calculate attention for the intersection operator \mathcal{I} . For community detection, we choose the Louvain algorithm [17]. To maintain comparability between conditions, we apply community detection *before* simulated edges are added to the original training data.

In terms of the training procedure, we set the learning rate to 0.001 and train the model with a batch size of 128 until convergence using Adam [41]. For every positive training sample, we randomly draw 10 negative samples. The negative sample set consists of randomly picked entities with the same type as the tail entity excluding all nodes that would satisfy the given relation; for example in the case where a user has retweeted more than one item.

We define various metrics that help us analyze network and community structure as well as quantify polarization between the communities. Concerning general graph statistics, we track average *degree* as well as network *density*. Since our goal is to reduce fragmentation, we also inspect the quality of community detection by calculating *modularity* [50] and *homophily* [46] to measure segregation. In order to explicitly account for *polarization*, we utilize

the metric proposed by Garimella et al. in [25]. Please note that we generalize it to a setting that includes more than two opposing communities. Finally, we track the average *acceptance probability* $P(\tau_r(u, t))$ of the respective integration procedure.

The knowledge graph embedding model has been implemented with PyTorch [56]. For several graph operations we use NetworkX [33].

5.1 Results

Our evaluation protocol is structured as follows: In a first step, we verify the effectiveness of our proposed approach in terms of link prediction quality. Afterwards, we present the results of the agent-based modeling procedure in detail.

5.1.1 Prediction Quality. Since our primary goal is not to improve the state-of-the-art in terms of accuracy and due to space limitations, we only conduct a small study in this regard. We split the original data into training (80%), validation (10%), and test set (10%). Please note that this part of the evaluation is therefore independent of edges simulated by agent-based modeling. In order to measure link prediction quality, we rank a positive sample against a set of 500 randomly selected negative samples. We report *hits@k* as well as *mean reciprocal rank* on the prediction of different edge types. We compare our method against a baseline variant that computes the dot product between head and tail entities without considering relational embeddings. Hence, the baseline represents conventional recommendation techniques, i.e. matrix factorization [42]. In case of multiple head entities, we calculate their average first.

The results, displayed in Table 1, show that our method consistently outperforms the baseline. It should be noted that, although the results for *Community* may appear comparably low, this behavior is expected as its main purpose is not the prediction of community membership for new users, but rather to learn an embedding located at the community center, thus representing an ideological position shared by many of its members.

Table 1: Prediction Quality on different edge types.

	Dot-Product Baseline			Knowledge Graph Embeddings		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
Tweet [$\mathcal{P}(t, \tau_t)$]	0.462	0.650	0.527	0.586	0.768	0.651
Retweet [$\mathcal{P}(u, \tau_r)$]	0.375	0.787	0.513	0.594	0.941	0.728
Community [$\mathcal{P}(c, \tau_c)$]	0.131	0.614	0.291	0.225	0.561	0.343
(User-)Recommendation [$\mathcal{P}(u, \tau_u)$]	0.657	0.922	0.766	0.716	0.974	0.817
Boundary-translation [$\mathcal{P}(\mathcal{P}(u, \tau_b), \tau_r)$]	0.308	0.781	0.460	0.585	0.959	0.717
Community-intersection [$\mathcal{I}(\dots)$]	0.512	0.865	0.684	0.668	0.986	0.798

5.1.2 Agent-based Modeling. *Degree* and *density* exhibit comparable characteristics in both the epistemic and ideological scenarios. In both cases, the *Ego-network* variant shows the lowest connectivity, while the other method, which refers to local information diffusion (*Recommendation*), reaches considerably higher levels. The difference between the two scenarios becomes clear when inspecting general graph statistics and the different evolution of the *Random* and the *Community-intersection* condition. While both show a comparatively high propagation of new connections in the epistemic scenario, this is visibly reduced in the ideological one according to our assumptions.

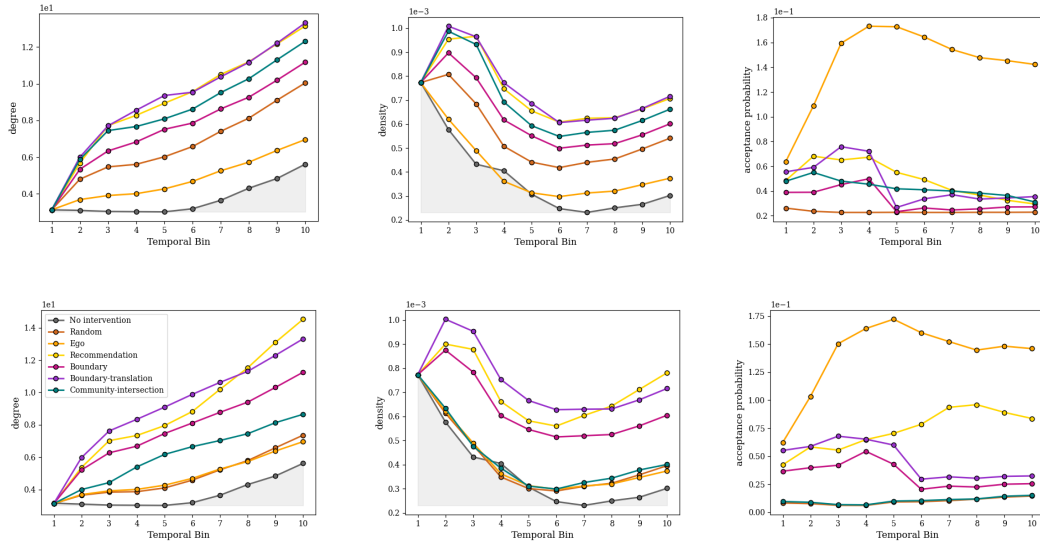


Figure 2: User-Graph statistics and acceptance probability for the epistemic (top) and the ideological scenario (bottom).

Comparing the average acceptance probability between both human filters shows that, in general, content is integrated with a higher rate in the epistemic condition. Additionally, as *Ego-network* and *Recommendation* yield the highest acceptance rates, it should be clear that people prefer close to more distant content. Note that, although the acceptance probability is highest for the *Ego-network* variant, this does not translate into higher connectivity in the user graph as content is primarily shared via already existing edges. Thereby, the *Recommendation* variant has a higher diversification potential because it is not limited to selecting candidate users based on graph connections. It follows that, over time, *Recommendation* may contribute to a widening of scope rendering it a more effective diversification strategy than connecting users across communities (*Random* and *Community-intersection*).

The *Random* baseline behaves notably different between the epistemic and the ideological scenario. In the epistemic case, more edges are added leading to higher connectivity and lower segregation (*modularity*, *homophily*, and *polarization*). In contrast, only small effects can be reported in the ideological scenario. The same applies to the *Community-intersection* condition: Notably, acceptance probability as well as depolarization capacity is almost identical to the *Random* condition. Therefore, we consider it not to be an effective strategy with respect to our operationalization of ideological filters.

The *Boundary* user intervention yields comparable results to the baselines in case of the epistemic scenario. Its benefit becomes apparent in the ideological scenario where it distinctly outperforms all baseline variants. Interestingly, the effectiveness is biggest during the first few interventions where we can observe a notable improvement in all measures as compared to the baselines. This is also true for the *Boundary-translation* condition. Yet, supporting our hypothesis, the selection of boundary users with respect to preferential dispositions seems to be even more effective. Another

finding worth noting is that, in both the *Boundary* as well as the *Boundary-translation* variant, there is a notable increase in *modularity* after time steps 4 (epistemic) and 5 (ideological) respectively. Interestingly, comparable developments cannot be observed in the other measures for segregation (*homophily* and *polarization*).

6 DISCUSSION AND CONCLUSION

Most scientific work that involves the simulation of opinion dynamics has treated social fragmentation in online spaces as the conjugate effect of biased information processing and a lack of coverage caused by unevenly distributed intra-network connectivity. What, in our opinion, is problematic about this conceptualization is the fact that through the missing distinction between both components the structure of cause and effect becomes opaque. For instance, [7, 15, 58] derive an individual's position in an attitude space by aggregating opinions from their local neighborhood. However, it remains unclear if biased information processing is the result of limited epistemic broadness or its cause [27]. What especially cannot be captured by respective models is the role of technology since interaction is modeled as a process of social influence. However, in reality, this process is to a significantly shaped by the communication channel.

Our findings underline this by showing that technological filters can have a large impact on the diffusion of information on social networks. At the same time, the effectiveness of diversification attempts is also dependent on the operationalization of individual human integration procedures. Interestingly, the degree of information diffusion is quite comparable between the *epistemic* and the *ideological* scenario if the technological filter restricts itself to local user neighborhoods (*Ego-network* and *Recommendation*). The distinction between both scenarios only becomes apparent with respect to more diversified recommendation sets. Specifically, we

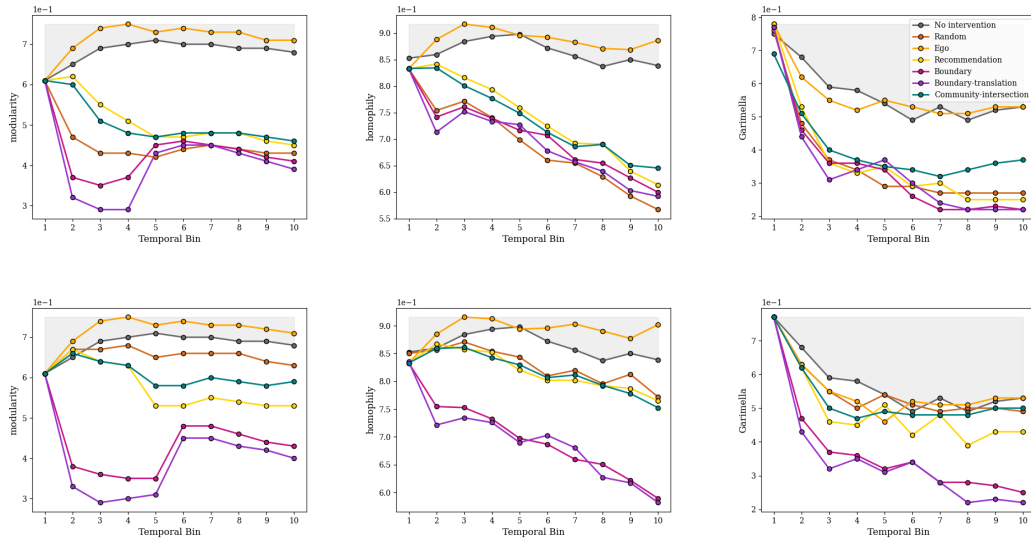


Figure 3: Segregation measures for the epistemic (top) and the ideological scenario (bottom).

find evidence that epistemic chambers might, indeed, not be as problematic as assumed. Even *randomly* selected users and tweets can help overcome chamber barriers. This finding is in line with hypotheses from literature stating that, in such cases, a mere lack of intra-network connectivity is responsible for epistemic impairments [55].

In case of pre-emptive distrust towards external sources, as represented by our *ideological* condition, bridging between communities becomes distinctly harder. Even the suggestion of outsider information with respect to personal preferences (*Community-intersection*) does not provide a perceptible increase in acceptance. This finding captures the concept, central for ideological echo chambers, that community insiders will be insulated from potentially valid counter-evidence streaming in from the outside.

Results regarding the *Boundary* and *Boundary-translation* conditions indicate that especially the border regions between communities can help to progressively open up the discussion. If witnessing the same argument by two different people, one from their own community, the other from another, the likelihood of acceptance is notably higher for the aligned source. It is, therefore, important to acknowledge the cognitive complexity that is involved in the socio-digital process. Settle shows in a series of lab experiments that social media usage increases the perceived differences between an individual’s own position and where they assume the out-group to be [59]. This suggests that the application of social identity schemata is fundamental to the processing of information in online environments.

Our study is not without limitations, though. Firstly, one valid concern is whether projections into the boundary region would cause backfire effects [52] instead of strengthening ties across communities. However, please note that we rely on retweets which rather indicate endorsement or at least critical commenting than opposing behavior [45]. We nonetheless agree that this is a relevant

question to address in the future. We admit that our model is not expressive enough to capture backfire effects adequately.

Furthermore, our agent-based modeling procedure has only been applied to a single data set. In order to fully evaluate the effectiveness of our proposed interventions, we see the need to analyze how the process shapes out in less polarized discussions as well. Additionally, we have solely focused on interactions observed on Twitter. The interaction dynamics on other social platforms may, however, be distinctly different [13, 70]. In future works, we intend to conduct a comparative study that involves varying degrees of polarization on different social platforms. In this regard, we also plan to verify the external validity of our method, for instance by conducting user studies designed to test the effectiveness of the proposed interventions.

Finally, one crucial aspect that remains unanswered by the study at hand concerns the question of how effective technological interventions can be in the first place as echo chambers are complex social phenomena that find their realization outside the digital domain as well. Nonetheless, our results shed some light on how technology might contribute to alleviation. It is particularly interesting that an intervention seems to be most effective when a discussion first emerges. This raises the question of why it does seem to be important what people observe when they first engage with a new topic. Under completely rational aspects, the order in which information is perceived should not matter. However, psychological research indicates that it does [35]. Thomas Kelly argues that if people would not spend too much weight on their initial sources, radical polarization would not play such an important role [40]. However, much cognitive effort is required to weigh different positions against each other. Therefore, people trapped inside an echo chamber tend to give too much importance to the evidence they encounter first. We follow from this that potential interventions should be applied as early as possible.

REFERENCES

- [1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.
- [2] Robert Axelrod. 1997. The dissemination of culture: A model with local convergence and global polarization. *Journal of conflict resolution* 41, 2 (1997), 203–226. Publisher: Sage Periodicals Press 2455 Teller Road, Thousand Oaks, CA 91320.
- [3] Mahsa Badami, Olfa Nasraoui, and Patrick Shafto. 2018. PrCP: Pre-recommendation Counter-Polarization.. In *KDIR*. 280–287.
- [4] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
- [5] Pablo Barberá. 2020. Social Media, Echo Chambers, and Political Polarization. *Social Media and Democracy: The State of the Field, Prospects for Reform* (2020), 34.
- [6] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26, 10 (2015), 1531–1542.
- [7] Fabian Baumann, Philipp Lorenz-Spree, Igor M Sokolov, and Michele Starnini. 2020. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters* 124, 4 (2020), 048301. Publisher: APS.
- [8] Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. *arXiv preprint arXiv:1406.3676* (2014).
- [9] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NIPS)*. 1–9.
- [10] Jennifer Brundidge. 2010. Encountering “difference” in the contemporary public sphere: The contribution of the Internet to the heterogeneity of political discussion networks. *Journal of Communication* 60, 4 (2010), 680–700. Publisher: Oxford University Press.
- [11] Axel Bruns. 2017. Echo chamber? What echo chamber? Reviewing the evidence. In *6th Biennial Future of Journalism Conference (FOJ17)*.
- [12] Axel Bruns. 2019. It’s not the technology, stupid: How the ‘Echo Chamber’ and ‘Filter Bubble’ metaphors have failed us. *International Association for Media and Communication Research* (2019).
- [13] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2020. Echo chambers on social media: A comparative analysis. *arXiv preprint arXiv:2004.09603* (2020).
- [14] Yuanfei Dai, Shiping Wang, Neal N Xiong, and Wenzhong Guo. 2020. A Survey on Knowledge Graph Embedding: Approaches, Applications and Benchmarks. *Electronics* 9, 5 (2020), 750.
- [15] Pranav Dandekar, Ashish Goel, and David T Lee. 2013. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5791–5796.
- [16] Rajarshi Das, Arvind Neelakantan, David Belanger, and Andrew McCallum. 2016. Chains of reasoning over entities, relations, and text using recurrent neural networks. *arXiv preprint arXiv:1607.01426* (2016).
- [17] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. 2011. Generalized louvain method for community detection in large networks. In *2011 11th international conference on intelligent systems design and applications*. IEEE, 88–93.
- [18] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. 2014. On Facebook, most ties are weak. *Commun. ACM* 57, 11 (2014), 78–84. Publisher: ACM New York, NY, USA.
- [19] Morris H DeGroot. 1974. Reaching a consensus. *J. Amer. Statist. Assoc.* 69, 345 (1974), 118–121.
- [20] Michela Del Vicario, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2017. Modeling confirmation bias and polarization. *Scientific reports* 7 (2017), 40391.
- [21] Elizabeth Dubois and Grant Blank. 2018. The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, communication & society* 21, 5 (2018), 729–745.
- [22] Leon Festinger. 1962. *A theory of cognitive dissonance*. Vol. 2. Stanford university press.
- [23] Dieter Frey. 1986. Recent research on selective exposure to information. *Advances in experimental social psychology* 19 (1986), 41–80. Publisher: Elsevier.
- [24] Noah E Friedkin. 2006. *A structural theory of social influence*. Number 13. Cambridge University Press.
- [25] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*. 913–922.
- [26] R Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of computer-mediated communication* 14, 2 (2009), 265–285. Publisher: Oxford University Press Oxford, UK.
- [27] Stefan Geiß, Melanie Magin, Pascal Jürgens, and Birgit Stark. 2021. Loopholes in the Echo Chambers: How the Echo Chamber Metaphor Oversimplifies the Effects of Information Gateways on Opinion Expression. *Digital Journalism* (2021), 1–27. Publisher: Taylor & Francis.
- [28] Daniel Geschke, Jan Lorenz, and Peter Holtz. 2019. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology* 58, 1 (2019), 129–149.
- [29] Salvatore Giorgi, Sharath Chandra Guntuku, Muhammad Rahman, McKenzie Himelein-Wachowiak, Amy Kwarteng, and Brenda Curtis. 2020. Twitter corpus of the #blacklivesmatter movement and counter protests: 2013 to 2020. *arXiv preprint arXiv:2009.00596* (2020).
- [30] Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia. 2014. People of opposing views can share common interests. In *Proceedings of the 23rd International Conference on World Wide Web*. 281–282.
- [31] Catherine Grevet, Loren G Terveen, and Eric Gilbert. 2014. Managing political differences in social media. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1400–1408.
- [32] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A Survey on Knowledge Graph-Based Recommender Systems. *arXiv preprint arXiv:2003.00911* (2020).
- [33] Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. *Exploring network structure, dynamics, and function using NetworkX*. Technical Report. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).
- [34] Will Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. 2018. Embedding logical queries on knowledge graphs. In *Advances in neural information processing systems*. 2026–2037.
- [35] Curtis P Haugtvedt and Duane T Wegener. 1994. Message order effects in persuasion: An attitude strength perspective. *Journal of consumer research* 21, 1 (1994), 205–218. Publisher: The University of Chicago Press.
- [36] Rainer Hegselmann, Ulrich Krause, and others. 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation* 5, 3 (2002).
- [37] John E Hunter, Jeffrey E Danes, and Stanley H Cohen. 2014. *Mathematical models of attitude change: Change in single attitudes and cognitive structure*. Vol. 1. Academic Press.
- [38] Kathleen Hall Jamieson and Joseph N Cappella. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.
- [39] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 383–390.
- [40] Thomas Kelly. 2005. The epistemic significance of disagreement. *Oxford studies in epistemology* 1, 167–196 (2005).
- [41] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [42] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37. Publisher: IEEE.
- [43] Hoon Lee, Nojin Kwak, and Scott W Campbell. 2015. Hearing the other side revisited: The joint workings of cross-cutting discussion and strong tie homogeneity in facilitating deliberative and participatory democracy. *Communication Research* 42, 4 (2015), 569–596. Publisher: SAGE Publications Sage CA: Los Angeles, CA.
- [44] Charles M Macal and Michael J North. 2005. Tutorial on agent-based modeling and simulation. In *Proceedings of the Winter Simulation Conference, 2005*. IEEE, 14–pp.
- [45] Anuja Majumdar, Jon-Patrick Allem, Tess Boley Cruz, and Jennifer Beth Unger. 2018. The why we retweet scale. *PLoS one* 13, 10 (2018), e0206076. Publisher: Public Library of Science San Francisco, CA USA.
- [46] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001), 415–444. Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.
- [47] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
- [48] Diana C Mutz. 2006. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press.
- [49] David G Myers and Helmut Lamm. 1976. The group polarization phenomenon. *Psychological bulletin* 83, 4 (1976), 602. Publisher: American Psychological Association.
- [50] Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103, 23 (2006), 8577–8582. Publisher: National Acad Sciences.
- [51] C Thi Nguyen. 2020. Echo chambers and epistemic bubbles. (2020).
- [52] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330. Publisher: Springer.
- [53] Enrico Palumbo, Diego Monti, Giuseppe Rizzo, Raphaël Troncy, and Elena Baralis. 2020. entity2rec: Property-specific knowledge graph embeddings for item recommendation. *Expert Systems with Applications* 151 (2020), 113235. Publisher: Elsevier.
- [54] Enrico Palumbo, Giuseppe Rizzo, Raphaël Troncy, Elena Baralis, Michele Osella, and Enrico Ferro. 2018. Knowledge graph embeddings with node2vec for item

- recommendation. In *European Semantic Web Conference*. Springer, 117–120.
- [55] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.
- [56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and others. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* (2019).
- [57] Hongyu Ren, Weihua Hu, and Jure Leskovec. 2020. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. *arXiv preprint arXiv:2002.05969* (2020).
- [58] Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2020. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science* (2020), 1–22. Publisher: Springer.
- [59] Jaime E Settle. 2018. *Frenemies: How social media polarizes America*. Cambridge University Press.
- [60] Muzafer Sherif and Carl I Hovland. 1961. Social judgment: Assimilation and contrast effects in communication and attitude change. (1961). Publisher: Yale Univer. Press.
- [61] Cass R Sunstein. 2001. *Republic. com*. Princeton university press.
- [62] Cass R Sunstein. 2018. *# Republic: Divided democracy in the age of social media*. Princeton University Press.
- [63] Henri Tajfel, John C Turner, William G Austin, and Stephen Worchel. 1979. An integrative theory of intergroup conflict. *Organizational identity: A reader* 56, 65 (1979), 9780203505984–16.
- [64] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018).
- [65] Xin Wang, Antonio D Sirianni, Shaoting Tang, Zhiming Zheng, and Feng Fu. 2020. Public discourse and social network echo chambers driven by socio-cognitive biases. *Physical Review X* 10, 4 (2020), 041042. Publisher: APS.
- [66] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28. Issue: 1.
- [67] David Westerman, Patric R Spence, and Brandon Van Der Heide. 2014. Social media as information source: Recency of updates and credibility of information. *Journal of computer-mediated communication* 19, 2 (2014), 171–183. Publisher: Oxford University Press Oxford, UK.
- [68] John Woods. 2005. Epistemic Bubbles.. In *We Will Show Them!(2)*. 731–774.
- [69] Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- [70] Moran Yarchi, Christian Baden, and Neta Kligler-Vilenchik. 2020. Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication* (2020), 1–42.
- [71] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 353–362.

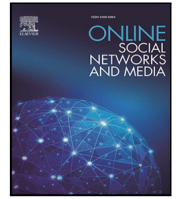
The following article is reused from:

Donkers, T., & Ziegler, J. (2023). De-sounding echo chambers: Simulation-based analysis of polarization dynamics in social networks. *Online Social Networks and Media*, 37–38, 100275.
<https://doi.org/10.1016/j.osnem.2023.100275>



Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

De-sounding echo chambers: Simulation-based analysis of polarization dynamics in social networks

Tim Donkers^{*}, Jürgen Ziegler

University of Duisburg–Essen, Forsthausweg 2, 47057 Duisburg, Germany

ARTICLE INFO

Keywords:

Social polarization
Echo chambers
Filter bubbles
Opinion dynamics
Social media
Machine learning
Latent space models
Recommender systems
Agent-based modeling
Knowledge-graph embedding

ABSTRACT

As online social networks have become dominant platforms for public discourse worldwide, there is growing anecdotal evidence of a concurrent rise in social antagonisms. Yet, while the increase in polarization is evident, the extent to which these digital communication ecosystems are driving this shift remains elusive. A dominant scholarly perspective suggests that digital social media lead to the compartmentalization of information channels, potentially culminating in the emergence of *echo chambers*. However, a growing body of empirical research suggests that the mechanisms influencing ideological demarcation are more complex than a complete communicative decoupling of user groups. This study introduces two intertwined principles that elucidate the dynamics of digital communication: First, socio-cognitive biases of social group formation enforce internal congruence of ideological communities by demarcation from outsiders. Second, algorithmic personalization of content contributes to ideological network formation by creating social redundancy, wherein the same individuals frequently interact in various roles, such as authors, recipients, or disseminators of messages, leading to a surplus of shared ideological fragments. Leveraging these insights, we pioneer a computational simulation model, integrating machine learning based on behavioral data and established recommendation technologies, to explore the evolution of social network structures in digital exchanges. Utilizing advanced methods in opinion dynamics, our model uniquely captures both the algorithmic delivery and the subsequent dissemination of messages by users. Our findings reveal that in ambiguous debate scenarios, the dual components of our model are essential to accurately capture the emergence of social polarization. Consequently, our model offers a forward-looking perspective on the evolution of network communication, facilitating nuanced comparisons with empirical graph benchmarks.

1. Introduction

The major digital networking platforms such as Facebook or Twitter have recently been the subject of much public criticism. Among other things, the increasing digitization of media and social exchange by the shift of entire communication systems to social platforms is held responsible for contributing significantly to social polarization both in the digital sphere and beyond [1–3]. To date, the causal mechanisms that underly social polarization have not been fully elucidated and understood. One explanation often given is that the personalized dissemination of platform content takes advantage of the human tendency to connect to like-minded people, with the main consequence then being the effective contraction of the available information pool [4–6]. It is now feared that this leads to disinhibited discourse shielded from dissonant opinion and criticism, which in turn bears the risk of fragmenting the formation of public opinion.

This phenomenon has been receiving scholarly attention under the term *echo chambers* for some time lately [7–12]. Anecdotal as well as empirical evidence indicates that communication networks, which often form spontaneously around certain topics, may indeed tend to spread centrifugally and, at the same time, condense communication processes. Contrary to all concerns, however, a complete delineation of discourse can by no means be observed [7,13]. Rather, several studies conclude that cross-community relationships, in terms of various engagement metrics such as likes, replies, and reposts, are prevalent in the digital domain, though the depth and nature of these interactions can vary significantly across platforms [8,14,15]. Moreover, the diversity of viewpoints users encounter on social networks is significantly greater than is the case, for example, with traditional dissemination media [16–18]. The assumption that the radicalization of certain groups takes

^{*} Corresponding author.

E-mail addresses: tim.donkers@uni-due.de (T. Donkers), juergen.ziegler@uni-due.de (J. Ziegler).

¹ While the notion that radicalization occurs in entirely isolated communication networks may not hold for major public platforms, it is crucial to recognize the potential influence of more insular spaces such as private Telegram or Facebook groups and partisan platforms.

<https://doi.org/10.1016/j.osnem.2023.100275>

Received 13 March 2023; Received in revised form 17 September 2023; Accepted 1 November 2023

Available online 30 November 2023

2468-6964/© 2024 Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

place in completely sealed off communication networks cannot be sustained, at least for the major platforms.¹

Nevertheless, the influence of digital social networks on the formation of opinion among the general public must be taken seriously as they serve as platforms where conflicts stemming from diverse interpretations and viewpoints are openly debated in front of expansive audiences [19]. In particular, emotionally or morally charged statements have the potential to grow into viral phenomena that are vehemently defended by supporters on the one hand and rigidly countered by oppositional voices on the other [20,21]. We argue that this inherent publicity of discourse basically turns the original echo chamber hypothesis on its head: social polarization is not the result of informational compartmentalization due to selective exposure, but precisely the fact that social media facilitates cross-cutting interaction moves people to sort themselves into ideological groups and, above all, distance themselves from other positions; predominantly by discrediting outsiders [13,22]. It is essential to note, however, that these perceptions can also be influenced by external narratives. Often, views about the 'out-group' are shaped by factors not present on the platform, with narratives being formed within the in-group, independent of direct interactions with the out-group. The nature of the discourse and the specific platform further modulate these dynamics.

Another aspect that must be taken into account when considering social polarization in the digital sphere concerns the fact that platforms span once-established chains of public communication, controlled by media corporations such as TV channels or newspaper publishers, and enrich them with new, egalitarian-seeming mechanisms of exchange. This *audience turn* allows more people to participate in wide-ranging communication, enabling former consumers of information to actively participate in the formation of public opinion. However, one major consequence of increasingly platform-dominated media environments is that traditional gatekeepers are being replaced by new distributors in the form of automated recommendation algorithms. Users oftentimes do not actively search for news anymore, but rather have it selected automatically in the form of personalized feeds without further manual intervention [23,24]. As such, personal horizons of information are largely subject to algorithmic control. The recommendation systems used are able to expand, condense or completely close off communication channels and may thus contain or drive social polarization. Problematically, as a rule, it is not possible to assess from the outside which parameter constellation was ultimately responsible for a certain content being displayed, because the underlying decision logic remains closed as a black box.

In summary, we consider both ideological demarcation and algorithmic dissemination as reciprocal contributors to social polarization in the digital public sphere. Following this line of thought, in this paper we present a computational model that draws from opinion dynamics literature as well state-of-the-art machine learning technology to simulate human communication and technological message dissemination in a unified framework. In doing so, we show how intensified identity formation via social groups becomes relevant to action in ambiguous settings and is consequently increasingly incorporated into the personalization of content.

Specifically, our simulation results indicate that ideological demarcation can in itself lead to progressive separation only if the initial debate space is to some degree pre-structured along ideological boundaries. However, especially in cases characterized by a high degree of ambiguity at the outset, the mere consideration of group identification is not sufficient to trigger polarizing effects. As we will show, a shared consensus is then only prevented if, additionally, algorithmic dissemination is personalized. On the one hand, this implies that the social organization of people with shared ideologies effectively depends on whether the platform's recommendation technology ensures that they find each other frequently enough. On the other hand, especially at the beginning a debate may be very divisive, so that ideologically dissonant messages need to be rigidly rejected in order to trigger corresponding effects on the dissemination technology and thus set a polarization spiral in motion.

2. Background

There exists a growing body of computational sociological and psychological research studying the relationship between digital media and social polarization. Influential approaches within opinion dynamics parameterize attitude formation and change primarily based on proximity relationships between agents via directly connected neighborhoods. In this model, individuals prefer to engage in homophilic exchange while being alienated with unfamiliarity or novelty [25–45]. This is justified by the notion that human sense-making processes are influenced by cognitive biases which manifest in people selectively exposing themselves to information that confirms their own world view, thus ensuring the maintenance of a cognitive equilibrium [46]. Although this reasoning is not entirely inaccurate, we will show that homophilic communication can only sustain a social divide for some time if certain network structures prevent the emergence of shared consensus through epistemic gaps.

However, since digital discourse spaces often drift apart despite (or even because of) cross-cutting communication, we argue that other psychological mechanisms besides selective exposure play a role. According to Jamieson and Cappella, echo chambers are the product of a systematic alienation of their members from external epistemic sources [47]. Dysfunctional behavioral patterns based on preemptive mistrust not only discourage people from going beyond their intellectual community in search of information, but more importantly, they also provide the impetus to actively discredit external voices; frequently by referring to conspiracy narratives. Where epistemic imbalances would imply the mere omission of opposing views, ideologically driven distrust can lead to collective foreclosure [10,48]. Thus, polarization often depends on the systematic exclusion of information sources by affective ideological demarcation from the outside [49]. Methodological approaches to bridging such boundaries by networking oppositional participants [e.g. 4,50,51] appear questionable against this background.

In fact, however, compartmentalization can not only be identified at the extreme edges of a debate. Rather, social differentiation between in-group and out-group seems to be such a fundamental element of social organization that comparable behavior is evident among representatives of mainstream viewpoints as well [52]. While their delimitation may be less openly communicated, respective socio-cognitive biases nonetheless contribute to the formation of certain actionable patterns of group dynamics [53–56]. From this perspective, the negotiation of ideological borderlines constantly takes place in social media discourse and oscillates, depending on the topic and societal structure, to a greater or lesser extent between (pluralistic) antagonisms and hegemonic dominance [57–59].

In consequence, we aim to extend the concept of local homophilic interaction patterns to incorporate the impact of effects of ideological demarcation. This helps us to sufficiently account for the far-reaching social influence that social groups and shared ideologies can have on attitude formation and subsequent communication decisions. Even if certain individuals are not directly connected to each other, mutual ideological structures that have grown out of socially acquired cognitions and are representative of a specific group identity may still exist. In this way, an ideological group marks an emergent self-referentiality that transcends the reductionist schema of focusing on individuals [13,48,60,61].

Furthermore, due to the conceptualization of communication as an exchange between directly connected neighbors, polarization is usually conceived largely as a social mechanism between actors known to each other. What is specifically crucial for digital platforms, however, is that the communication process is structured to a significant degree by the underlying dissemination technology. In our view, the distribution of messages via direct connections alone does only insufficiently take this technological influence into account. In many models it is blanketly assumed that an agent is confronted with generic messages emigrating

from their close-tie communication network, and subject a certain probability, an adjustment of their position in the attitude space takes place as a result. However, horizons of information generated by recommendation technology thus remain vague. In reality, social platforms use sophisticated algorithms that consider numerous parameters to make personalized content selections [62].

Summarized, in order to simulatively capture social network polarization, we require a model that adequately approximates the reciprocal relationship between human communication decisions on the one hand and dissemination technology on the other. To this end, we aim to exploit the structural properties of (empirical or synthetic) social graphs by transferring the relations between human communication participants and messages distributed through technology into a shared semantic information space. In contrast to established approaches, we do not explicitly position actors in an attitude space, however. Instead, we use a machine learning technology based on knowledge graph embeddings [63] that allows us to extract and extrapolate latent relationships from an existing graph topology. The learned representations of entities can then be related to each other [cf.64,65] to determine, for example, the probability with which a user will redistribute an incoming message. For the case of dissemination technology, it has already been sufficiently shown that related methods are capable of generating high-quality recommendations in different contexts [cf.62,63,66].

3. Simulation agents

The latent space model we want to present in this paper is essentially based on the assumption that generalizable structural relationships can be derived between the different nodes of a social graph (see Appendix A for a detailed derivation of the model). We do not merely consider social connections between actors, however, but intermediate their relationship through the messages exchanged over the network. The resulting heterogeneous communication graph \mathcal{G}_c is, hence, comprised of a set of user nodes \mathcal{V}_u and message nodes \mathcal{V}_m . Users are initially connected to messages via two types of edges: First, messages can be composed by them (τ_c), second, they can react to incoming messages in the form of a redistribution (τ_r). A well-known example of the latter is Twitter's retweet functionality. This allows us, for example, to specifically capture which messages a user has reacted to in order to estimate their personal preferences.

Building upon the observed topological correlations inherent in the graph structure, our working hypothesis posits that these patterns can be effectively modeled and generalized by operationalizing the semantics of a latent space. This model is learned based on the structural information of the graph, capturing the intricate relationships between nodes and the different types of edges.

It is important to note that when we refer to semantics in the context of this work, we are not alluding to linguistic meanings typically associated with natural language processing. Instead, we use the term to describe the inherent patterns and relationships captured by the latent space models from behavioral graph data. This perspective aligns with the broader understanding of semantics in cognitive science, where semantic networks have been used to represent knowledge structures, with nodes representing concepts and edges denoting relationships between them [67].

Pioneering work on the quantitative representation of semantic contexts has shown that appropriately trained latent spaces organize semantically similar entities in close proximity to each other [cf.64]. We want to harness these latent semantics by learning geometric operators together with the latent representation of graph nodes. Detailed explanations of such geometric operators can be found in [48,65]. For the present case, we will restrict ourselves to a projection operator \mathcal{P} , which translates a given latent representation $\mathbf{q} \in \mathbb{R}^d$ in a space of dimensionality d subject to a certain edge type τ with representation $\mathbf{r}_\tau \in \mathbb{R}^d$:

$$\mathcal{P}(\mathbf{q}, \tau) = \mathbf{q} + \mathbf{r}_\tau \quad (1)$$

For instance, we may apply geometric operator $\mathcal{P}(\mathbf{u}, \tau_r)$ to user node $u \in \mathcal{V}_u$ to identify a region in information space where message nodes are in close proximity to each other that this user has a high probability of propagating. Note that \mathbf{q} can represent the original latent representation of a concrete graph node as well as one that has already been shifted one or more times. For example, $\mathcal{P}(\mathcal{P}(\mathbf{u}, \tau_r), \tau_s)$ searches for actual linked users as well as potentially linkable matches for a target user u .

A projected latent representation can subsequently be interpreted geometrically by examining the distances to graph nodes of a certain type. For this we introduce δ as a second operator which specifies the normalized Euclidean distance between a projected representation \mathbf{q} and a target node \mathbf{v} :

$$\delta(\mathbf{q}, \mathbf{v}) = \frac{1}{2} \frac{\text{Var}(\mathbf{q} - \mathbf{v})}{\text{Var}(\mathbf{q}) + \text{Var}(\mathbf{v})}, \quad (2)$$

where Var is the variance of the input vector and $\mathbf{v} \in \mathbb{R}^d$ is the vector representation of \mathbf{v} .

For the combined application of one or more projection operations to an anchor node v_1 with a final calculation of distance to a target node v_2 , we write $\delta_p(v_1, v_2 | T)$, where $T = (\tau_1, \dots, \tau_n)$ is a tuple of projection edges. Starting from this central operation, we now aim to explore how we can utilize the latent structures of the information space, derived purely from graph topology, to relate the entities of a social graph and develop a corresponding simulation procedure. Specifically, we want to take the geometric relationships between entities as a basis for an opinion dynamics model that determines whether or not a user will propagate a particular message.

As already indicated earlier, however, we do not assume that messages are exclusively socially disseminated in the form of broadcasting to an immediate neighborhood. While humans still have primacy in our framework over whether to respond to an incoming message, a recommender system autonomously decides which messages human participants get to see in the first place. Potentially, then, a message can be played out beyond the boundaries of a neighborhood or only to a subset of it. In contrast to established approaches, this gives the technological component the rank of a full-fledged agent. It stands in a reciprocal exchange with the users insofar as it, for one, selects the messages for each human participant individually — thereby allowing us to get a grasp of the phenomenon of filter bubbles [e.g.6,68] as well. Conversely, user decisions to react or not to a recommended message are fed back to the system, influencing future recommendation generation.

Put together, in the following, we will first describe both types of agents, human and technological, individually before bringing them together in a shared framework in the next section.

3.1. Human agents

Suppose the situation of a user, faced with a novel input message, who needs to decide whether or not to react to it in some way, for instance by further propagating it. To model this, we refer to advances in agent-based modeling (ABM) [27]. The core idea behind ABM is that the actions of certain agents, for instance humans, can be modeled by applying a mathematical decision function simulating a reaction towards incoming new input. In our case, this decision function is inspired by psychological behavioral research.

When discussing the psychological basis of social polarization as well as the formation of echo chambers in the literature, selective exposure to content, or confirmation bias, is often cited as an explanation [9,11,12]. Therefore, we will now first present a decision function inspired by this kind of individual demarcation. Later, we will propose a second function extending the first by additionally considering ideological demarcation patterns.

3.1.1. Individual demarcation

The sense-making processes that are triggered during the evaluation of new messages are initially attempts to classify what is perceived into a personally meaningful structure. This individual frame of reference implies that the generation of meaning is anchored in the construction of one's own identity [69]. However, sense-making processes require quite a range of cognitive heuristics, mental shortcuts so to speak, to function efficiently. Important to our case is the human tendency to prefer things they are well acquainted with which implies that information, that contradicts one's internalized assumption about a phenomenon or that cannot easily be integrated into one's belief system, often tends to be disregarded or ignored [e.g.46].

Translated to latent space semantics, we assume that the likelihood of connecting communication is high if a new message is structurally similar to messages with which the user has interacted in the past. We take inspiration from social judgment theory [70] to model this with respect to a concept called *latitude of acceptance*. Whenever a new bit of information comes to the attention of an individual, we apply a bounded confidence model to determine the likelihood of connecting communication as:

$$\mathcal{L}_1(\tau_r(u, m)) = \frac{\lambda^\mu}{\delta_p(u, m | \tau_r)^\mu + \lambda^\mu}, \quad (3)$$

where $u \in \mathcal{V}_u$ and $m \in \mathcal{V}_m$ are user and message nodes respectively, $\lambda \in [0, 1]$ is the latitude of acceptance, and $\mu \in \mathbb{N}$ is a sharpness parameter that determines how steeply the likelihood of connecting communication drops from one to zero around the latitude of acceptance. Informally, a user representation is projected into the area in latent space where messages with already established edges are located. Then, the distance to the target message is calculated and fed into a non-linear decision function.

3.1.2. Ideological demarcation

As we have previously described, we view the cultivation of identity through differentiation from other social groups as an essential feature of social organization in the digital sphere. At the level of communication, this group behavior manifests itself in an affective delimitation from outsiders, which is characterized by internalized antipathy and pre-emptive mistrust. The internal cohesion of such a social group is, hence, primarily ensured by its external relationships, especially towards other social groups. In this sense, the formation of a shared group ideology depends to a large extent on whether a social group achieves a communication-relevant differentiation between in- and out-group.

To represent the identification with a group, we introduce a new type of node, namely communities, as $c \in \mathcal{V}_c$. Users being members of a community is expressed via edge type τ_c . The method by which these edges are derived inductively is in principle arbitrary. Nonetheless, some considerations have to be made. To get an idea of the social structure embedded in the communication graph \mathcal{G}_c , we explicate the connections that can be derived from it to build a social graph \mathcal{G}_u . Tracing paths along both τ_r and τ_s as a conjunction of edges allows us to query \mathcal{G}_c in order to identify connected users. We assume a directed edge from user u_1 to user u_2 to exist if u_1 has engaged in connecting communication with respect to a message m authored by u_2 . Please note that edges between users are, thus, not observed directly (e.g. via explicit following relationships) but are indirectly derived from communication. This is indeed an important distinction since follow relations often say little about one's social identity. Several studies indicate that subscribing to people who belong to different ideological camps happens frequently [8,14,15]. In contrast, we interpret onward dissemination of content as an expression of assumed relevance for one's own social group.

Based on \mathcal{G}_u , we perform community detection to derive the membership edges. We want to emphasize that we assume group identity behavior to only be derivable from acts of connecting communication, i. e. social support expressed by reacting to a message. Opposed to this,

sending a message initially does not tell us anything about the social relations or intentions of an author. Therefore, we utilize a directed variant of Leiden's community detection algorithm [71] such that users are assumed to form a community if they react similarly to a given set of messages.

After determining community membership, we transfer these new edges into the latent information space by training a projection $\mathcal{P}(u, \tau_c)$. Unlike the original community detection, where membership is determined in a binary fashion, projection allows us to locate it on a spectrum of real-valued numbers. Informally speaking, the distance $\delta_p(u, c | \tau_c)$ of projected user u to community c , in terms of the previously described semantics of the information space, depicts how strongly a user represents the shared characteristics of a community.

We now want to translate this relationship to our simulation procedure by taking the distance as a measure of how strongly a user identifies with a group. Following prototype theory, we, hence, describe membership of entities to certain categories by a quantitative gradation [72]. Prototypes are central normative category elements to which other entities can only be determined relatively with a more or less large distance. Thus, prototype semantics assume that ideal categorizations never take place exactly, but can only gradually approximate an abstract prototype. Thereby, we calculate the strength of group influence $\eta \in [0, 1]$ by the relation between a user's prototypicality with respect to in- and out-group:

$$\eta = \frac{2\delta_p(u, c_u | \tau_c)^\kappa}{\delta_p(u, c_u | \tau_c)^\kappa + \delta_p(u, c_m | \tau_c)^\kappa}, \quad (4)$$

where $\kappa \in \mathbb{N}$ is an ideological sharpness parameter and c_m is the representation vector of the community from which a message m originates.

We let this conceptualization of group identity influence communication by shifting the latitude of acceptance depending on whether a new input originates from the in- or an out-group. Formally, we define group influence as a weight on the latitude of acceptance:

$$\mathcal{L}_2(\tau_r(u, m)) = \frac{\eta\lambda^\mu}{\delta_p(u, m | \tau_r)^\mu + \eta\lambda^\mu} \quad (5)$$

The influence becomes stronger the more an individual identifies with their own group's views and the less they identify with an out-group from which the input originates. Note that this formula completely abstracts from the author of a message. Instead, messages are viewed as being representative of their specific social group. This is a conscious modeling decision to carry on with the idea that out-group members are met with preemptive mistrust. Finally, also consider the fact that $\eta = 1$ in case a new input message stems from a user's in-group. As a result, for in-group communication \mathcal{L}_2 is equivalent to \mathcal{L}_1 .

3.1.3. Rewiring

On a final note, some works have pointed out the importance of dissolving existing links to model the conscious decision of users to cut ties with each other and thus potentially the transition into another ideological sphere. [e.g.41]. Accordingly, we propose a method to implement disconnecting behavior with respect to our framework. Concretely, we randomly cut links to $x\%$ of the existing message edges and their respective authors while making sure that no user gets isolated. Internal pre-tests indicate that certain effects occur earlier in our model if we follow a rewiring strategy. Notice, however, that the principal development of communication remains unaffected by this.

3.2. Technological agent

In order to approximate the dissemination technology underlying social networks, we define a recommendation task where a model selects a number of top- n recommendations from a large catalogue of target entities subject to a specific decision rule. In our case, the recommendation process is split into two phases. Since we are mainly interested in how user relations evolve in a given scenario, we first identify users to connect. This is in accordance to modern social media

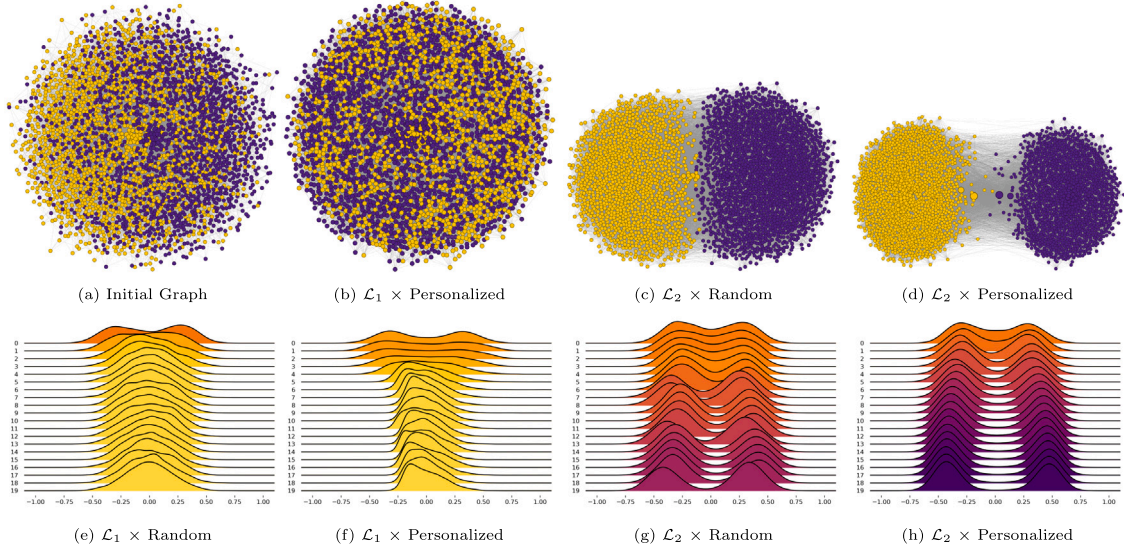


Fig. 1. Exemplary visualizations of graph topologies and latent space polarization for the partisan scenario. The graphs shown for each condition (Fig. 1(b) to 1(d)) are taken from one randomly selected simulation run at final epoch 100. While in the initial graph (Fig. 1(a)), the two communities are strongly intertwined, all conditions using ideological demarcation are able to clearly separate them, with $\mathcal{L}_2 \times$ Personalized achieving the strongest delineation. Condition $\mathcal{L}_1 \times$ Random is omitted because the graphs shows no notable effects. Below (Fig. 1(e) to 1(h)), polarization in latent space is depicted in the form of joyplots (also known as ridgeline plots) giving a sense of the evolution of polarization over time (y-axis). The plots show the kernel density estimations of a one-dimensional principal component analysis (PCA) on the latent representations of user nodes. The color gradient from yellow (low) to violet (high) depicts the PCA’s explained variance. The results reveal that the slight separation at the start cannot be maintained by conditions $\mathcal{L}_1 \times$ Random and $\mathcal{L}_1 \times$ Personalized, while the remainder increase bipolarization, with $\mathcal{L}_2 \times$ Personalized showing the most pronounced effects.

platforms in that the content displayed to users in news feeds stems to a large degree (but not exclusively) from people who are directly connected to a target user, i.e. the messages are drawn from a user’s follow network. This implies that platforms generally do not only apply filtering methods to the content to recommend alone, but rather they pre-select a set of users from which a potentially relevant set of messages is then selected afterwards.

For both steps, we exploit the learned model to formulate a context-aware ranking problem where the target entities are either users or messages. By context, we specifically mean the individual user for who recommendations are to be generated. The goal of entity recommendation is, hence, to retrieve a subset of entities for a given user $u \in \mathcal{V}_u$ from a set of target entities $\mathcal{V}_t \subset \mathcal{V}$. The preference of u to a target entity $t \in \mathcal{V}_t$ is given by a scoring function \hat{y} :

$$\hat{y}(u, t) = \hat{y}(t | u), \hat{y} : \mathcal{V}_u \times \mathcal{V}_t \rightarrow \mathbb{R} \quad (6)$$

This scoring function is parametrized by a set of model parameters which we draw from our previously introduced latent space model. It can be applied to rank target entities for a given user. Let $\hat{r}(t | u)$ be the rank (or position) of target t in the sorted list of entities given a user u and a scoring function \hat{y} , i.e. $\hat{r} : \mathcal{V}_t \rightarrow \{1, \dots, |\mathcal{V}_t|\}$. Here, \hat{r} is a bijective mapping such that the inverse function gives the entity ranked at a certain position. When a recommendation model \hat{y} is used for the task of target recommendation, it needs to find the n highest scoring entities for u . That means it has to compute the items $\hat{r}^{-1}(1 | u), \dots, \hat{r}^{-1}(n | u)$.

For now, we will restrict ourselves to proposing only very basic implementations of both user and item recommendation. In order to identify the set of message candidates C_m in the first place, we rank the set of target users first. We utilize graph queries to define a scoring function $\hat{y}_u : \mathcal{V}_u \times \mathcal{V}_u \rightarrow \mathbb{R}$ as:

$$\begin{aligned} \hat{y}_u(u, v) &= \hat{y}_m(v | u) \\ &= \delta_p(u, v | (\tau_r, \tau_s)) \end{aligned} \quad (7)$$

Informally, this query first asks where in latent space relevant messages are found, to then, based on the identified position, move on to the location of author users that are likely to have posted messages with the respective structure. In other words, it tries to identify both already existing connected users and potential new matches. The subsequent

ranking $\hat{r}_u : \mathcal{V}_u \rightarrow \{1, \dots, |\mathcal{V}_u|\}$ gives the required set of users $C_u = \{\hat{r}_u^{-1}(1 | u), \dots, \hat{r}_u^{-1}(n | u)\}$. From C_u , we can derive the set of message candidates C_m as:

$$C_m = \{m | \exists U : \tau_s(m, U) = 1 \wedge U \in C_u\} \quad (8)$$

The items in C_m can be ranked via $\hat{r}_m : C_m \rightarrow \{1, \dots, |C_m|\}$ subject to another scoring function \hat{y}_m to retrieve the set of messages to display to u as $\{\hat{r}_m^{-1}(1 | u), \dots, \hat{r}_m^{-1}(n | u)\}$:

$$\begin{aligned} \hat{y}_m(u, m) &= \hat{y}_m(m | u), m \in C_m \\ &= \delta_p(u, m | \tau_r) \end{aligned} \quad (9)$$

Please note that, naturally, we exclude all messages from C_m towards which the target user has already established a link.

4. Simulation-based validation

Having described all the components of our model in the previous section, we now want to conceptualize simulation procedures to hypothesize and play out a range of different scenarios.² Basically, we start from a synthetically generated or an observed real-world communication graph to train our proposed prediction model against the corresponding loss function in Eq. (A.4) until convergence. Afterwards, we simulate the evolution of network communication over a period of $T = 100$ epochs, splitting the procedure into two key steps: First, a set of message recommendations is generated for each user. Following this, secondly, the process of connecting communication is modeled with respect to the messages recommended. Note that the choice of the decision function (\mathcal{L}_1 vs. \mathcal{L}_2) depends on the predefined scenario. In this context, exclusively self-referential or group-referential scenarios as well as combinations of both are conceivable. The number of identified communities is also variable, so that not only two-sided polarization effects can be modeled, but any number of social groups can act competitively against each other.

² The code can be accessed via <https://github.com/ti-ra-do/de-sounding-echo-chambers.git>.

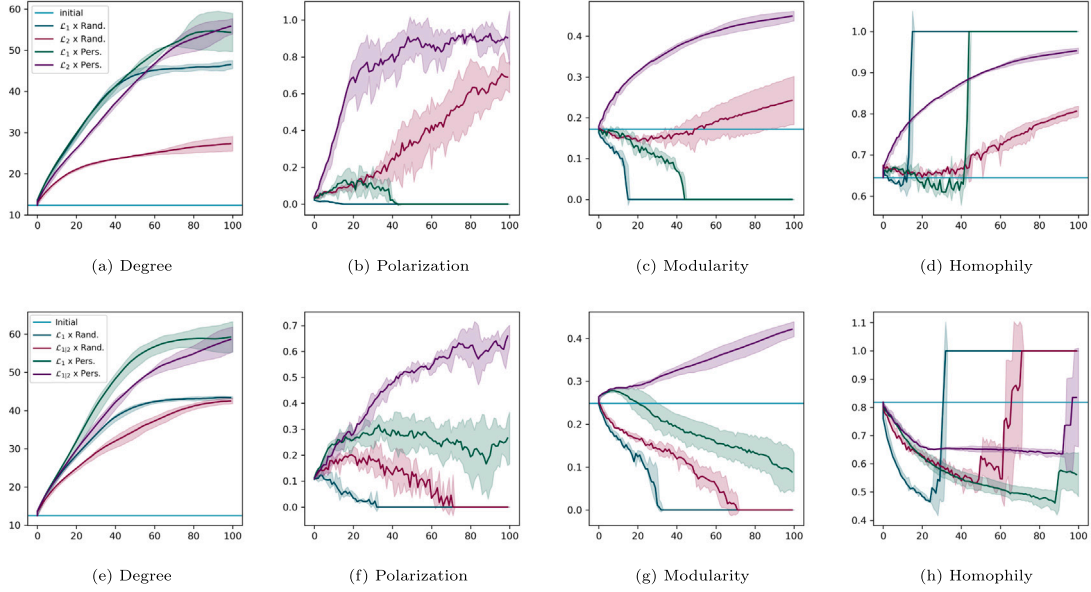


Fig. 2. Metrics for both the partisan (Fig. 2(a) to 2(d)) as well as the divisive scenario (Fig. 2(e) to 2(h)). Results depict the average of 5 runs for each condition with the shaded area depicting the standard deviation. The blue straight line represents values for the initial graph.

In our simulations, when a system recommendation $r^{-1}(n|u)$ is accepted by user u , it produces a new edge $\tau_r(u, r^{-1}(n|u))$. These newly formed edges influence the subsequent simulation epoch, refining the predictive model to adapt to the evolving network structure. Communities within the network are recalculated dynamically in each epoch using the directed variant of the Leiden algorithm, as detailed in Section 3. Given the fluid nature of communities in social networks, we use a reference-based approach to maintain their identity. Communities detected in the previous epoch serve as this reference. For every community identified in the current epoch, we compute the Jaccard index [73] to assess its overlap with communities from the reference. This approach allows for the natural evolution of communities, accommodating the potential migration of users between different communities.

In order to measure network polarization, we apply homophily [74] and modularity [75], both of which refer to the existing communities' intra- and inter-connectivity. Additionally, we observe polarization in latent space by tracing the distances between groups as:

$$\Psi = \frac{1}{|\mathcal{V}_u|} \sum_{u \in \mathcal{V}_u} \Delta_{ex}(u) - \Delta_{in}(u), \quad (10)$$

where $\Delta_{in}(u_1) = \frac{1}{|\mathcal{V}_{c_u}|} \sum_{u_2 \in \mathcal{V}_{c_u}} \delta_p(u_1, u_2 | (\tau_r, \tau_s))$ is the average of internal distances for community c_u of user u and $\Delta_{ex}(u_1) = \frac{1}{|\mathcal{V}_u \setminus \mathcal{V}_{c_u}|} \sum_{u_2 \in \mathcal{V}_u \setminus \mathcal{V}_{c_u}} \delta_p(u_1, u_2 | (\tau_r, \tau_s))$ is the average of external distances towards users in the remaining communities respectively.

4.1. Scenarios

To validate our proposed framework we aim to show, first, that the choice of human decision function has a significant impact on whether a social group stabilizes over time or collapses. In addition, we want to shed light on the relationship between human communication and recommendation algorithm selection. In particular, we are interested in investigating whether the personalization logic underlying social media recommendations can contribute to group stability and possibly even fuel centrifugal tendencies.

The primary human decision functions we consider are the previously introduced decision-functions \mathcal{L}_1 and \mathcal{L}_2 . In our study, we also recognize scenarios where communities exhibit varying degrees of ideological demarcation. Specifically, while some communities clearly

align with distinct ideological stances (guided by \mathcal{L}_2), others remain in a state of flux, with individuals not yet committing to a clear ideological position (guided by \mathcal{L}_1). To model such heterogeneous scenarios, we introduce a hybrid decision function, denoted as $\mathcal{L}_{1|2}$. This function captures situations where individuals within different communities might either demarcate themselves based on their personal inclinations or look to their group as a reference for alignment.

Regarding the recommendation algorithms, the main independent variable is $\mathcal{A} \in \{\text{personalized}, \text{random}\}$. By ‘random’, we mean that message recommendations are not calculated with respect to some personalization logic, but are rather chosen arbitrarily from any graph user. This yields a two-fold $\mathcal{L} \times \mathcal{A}$ study setup, with $\mathcal{L} \in \{\mathcal{L}_1, \mathcal{L}_2\}$ or $\mathcal{L} \in \{\mathcal{L}_1, \mathcal{L}_{1|2}\}$ depending on the specific scenario.

We apply this setup to synthetic communication graphs with different levels of group integration and, accordingly, degrees of polarization below. See Appendix B for a detailed explanation of our method to create synthetic graphs. Further analysis regarding the variation in community-internal and external connectivity as well as the relative community sizes can be found in Appendix D. Finally, we present preliminary results regarding the application of our model to empirical data in Fig. 4 as well as Appendix F.

To explore our ostensible research question about the relationship between ideological demarcation and algorithmic dissemination, we will focus here on two scenarios: First, a partisan case in which personal identification with one of two camps is predetermined at the outset — although subsequent migrations are not ruled out. In the second scenario, we describe an unaligned case, in which two points of view diverging from the mainstream have gathered at the fringes of the debate, while the bulk of the participants have not yet taken a fixed position. See Figs. 1(a) and 3(a) for an overview of the initial graph topologies for each scenario.

4.1.1. Scenario 1 — partisan case

In our first setup, we aim to model the process of two communities competing against each other. We assume both communities to exhibit some degree of homophily while still being intertwined. Formally, we define two communities $\mathcal{V}_c = \{c_1, c_2\}$ with $|\mathcal{V}_u[c_1]| = |\mathcal{V}_u[c_2]| = 2,500$ (100 gatekeepers each). For each user, we set their connection probability to $\mathbf{p}[c_1] = (1 - \alpha_u, \alpha_u)$ and $\mathbf{p}[c_2] = (\alpha_u, 1 - \alpha_u)$ respectively with $\alpha_u = 0.4$ being the external connectivity. Moreover, gatekeepers

are to be modeled as overly active users, so they are expected to be connected to on average 40 other users, which we select according to the method suggested in Appendix B. For the remaining users, we assume 5 connections. The user graph consists of 31,498 edges. For the communication graph, we append 39,042 message nodes and observe 169,478 edges.

4.1.2. Scenario 2 — unaligned case

In the second scenario, we want to inspect the case where a large proportion of the population is not yet leaning towards a specific standpoint. Oftentimes, especially in case a topic is highly ambiguous, many people are undecided with which side of a debate to solidarize while others quickly adopt a standpoint that aligns with the view of their ideological group. Hence, we propose the community of undecided interlocutors to follow \mathcal{L}_1 until they have decided for a standpoint while members of the two fringe communities apply \mathcal{L}_2 , i.e., the hybrid decision function $\mathcal{L}_{1|2}$ is applied.

Formally, we define three communities $\mathcal{V}_c = \{c_0, c_1, c_2\}$ where $\mathcal{L}_{c_0} = \mathcal{L}_1$ and $\mathcal{L}_{c_1} = \mathcal{L}_{c_2} = \mathcal{L}_2$. We set $|\mathcal{V}_u[c_0]| = 4,000$ (with 100 gatekeepers) and $|\mathcal{V}_u[c_1]| = |\mathcal{V}_u[c_2]| = 500$ (with 10 gatekeepers each). Concerning inter-community connectivity, we let c_1 and c_2 be entirely separated from each other. Still, both produce some edges towards the unaligned community c_0 and vice versa. Accordingly, $\mathbf{p}[c_0] = (1 - 2\alpha_u, \alpha_u, \alpha_u)$ as well as $\mathbf{p}[c_1] = (\alpha_u, 1 - \alpha_u, 0)$ and $\mathbf{p}[c_2] = (\alpha_u, 0, 1 - \alpha_u)$ with $\alpha_u = 0.1$. The resulting user graph consists of 32,528 edges and the communication graph has 38,803 message nodes and a total of 169,096 edges.

4.1.3. Hyperparameters

Considering relevant hyperparameters, various pre-tests have shown that our method can robustly generate the intended effects for different constellations. We obtained the most stable results with respect to latitude of acceptance in the range $[0.3, \dots, 0.7]$ (compare Appendix E). For the present study, we decided to set $\lambda = 0.5$. Furthermore, we fix both the regular as well as the ideological sharpness parameter to $\mu = \kappa = 5$ and the disconnect ratio to 10%. Finally, the number of message recommendations is 20 per epoch per user. Reasonable values for further hyperparameters, especially with respect to the machine learning method, can be taken from our Github repository.

4.2. Results

The results for Scenario 1 indicate a continuous increase in edge connectivity for all conditions (Fig. 2(a)). The slope decreases with time and thus a stable state of the graph degrees is soon established. The reason for this is that in earlier episodes proportionally more new connections are made than are cut. However, when a certain connectivity level is reached at some point, the ratio of new proposed and accepted users to disconnecting old connection equalizes. Most edges are, thereby, formed in the two conditions that follow decision rule \mathcal{L}_1 . While condition $\mathcal{L}_2 \times \text{Personalized}$ has a comparable, only slightly lower degree, the level is significantly lower in the case of $\mathcal{L}_2 \times \text{Random}$.

Considering the measures of network polarization and segmentation (Fig. 2(b) to 2(d)), we essentially observe the same picture for all three. The random distribution of messages under the condition of only homophilic interaction ($\mathcal{L}_1 \times \text{Random}$) exhibits a very low level of polarization. Modularity and homophily decrease noticeably from the starting point of the initial graph and then remain stable at low levels. Polarization of latent representations is also small. For the case of personalized message dissemination ($\mathcal{L}_1 \times \text{Personalized}$), we note that the debate space can maintain separation at least for a while. Eventually, however, it converges reliably as well. Elevated levels of polarization are only achieved if we exchange the decision rule and

apply \mathcal{L}_2 . Thus, under the given network structure the debate space fragments only if appropriate group behavior is exhibited, whereby this effect is most evident when a personalized dissemination strategy is pursued ($\mathcal{L}_2 \times \text{Personalized}$).

Overall, these effects occur quite robustly over several simulation runs. Even in the latent information space, where smaller fluctuations occur, the described relationships are stable (Fig. 2(b)). If we plot dimensionally reduced representations of the user nodes in the latent information space over time in the form of joyplots, the drifting apart or consensual coming together of the communication participants can be easily followed (Fig. 1(e) to 1(h)). While the separation in the latent information space is readily apparent at the beginning, general agreement quickly emerges for $\mathcal{L}_1 \times \text{Random}$ and some time later for $\mathcal{L}_1 \times \text{Personalized}$. For the remaining two conditions, the Probability Density Estimation drifts apart, with $\mathcal{L}_2 \times \text{Personalized}$ separating the two social groups very clearly.

The described fragmentation of the debate space can also be traced in the graph topology (Fig. 1(b) to 1(d)). In the initial graph, slight condensation tendencies can be observed, but overall the distribution of the nodes is very mixed. For the presented conditions, however, the picture of latent representation is mirrored at the graph level.

Turning now to Scenario 2, we first note that, unlike in the first, $\mathcal{L}_1 \times \text{Random}$ settles at a lower level than $\mathcal{L}_1 \times \text{Personalized}$ and $\mathcal{L}_{1|2} \times \text{Personalized}$ (Fig. 2(e)). The evolution of the other conditions is comparable to before. With respect to the polarization measures (Fig. 2(f) to 2(h)), we document higher fluctuations between simulation runs, but nevertheless the direction of the effects is clear. $\mathcal{L}_1 \times \text{Random}$ is hardly able to maintain or even advance the initial low separation as the debate space quickly collapses consensually. We observe similar effects for the case $\mathcal{L}_{1|2} \times \text{Random}$, although the separation of the communities is maintained for longer. Eventually, however, they still converge. The network graph of $\mathcal{L}_1 \times \text{Personalized}$ becomes less modular over time, but latent representations are polarized at least at a low level. The only condition, however, that forms clearly discernible polarization patterns both in latent space and at the graph level is $\mathcal{L}_{1|2} \times \text{Personalized}$.

In addition, it should be noted that the measure of homophily (Fig. 2(h)) has some conspicuous features. It decreases initially across all conditions since the starting graph is already notably homophilic due to the proportional dominance of the unaligned community. However, for the conditions $\mathcal{L}_1 \times \text{Personalized}$ and $\mathcal{L}_{1|2} \times \text{Random}$ respectively, sudden jumps to a value of 1.0 can then be observed, which is simply due to the three communities collapsing into one. Comparable jumps can also be observed for the other two conditions, which, however, settle at values lower than 1.0. Again, the collapse of communities is responsible; however, not into one, but into two. It is important to note that $\mathcal{L}_{1|2} \times \text{Personalized}$ produces significantly more homophilic communities than $\mathcal{L}_1 \times \text{Personalized}$, however.

This evolution of membership can actually be better traced in the latent information space (Fig. 3(e) to 3(h)). If we transfer the distributions to a two-dimensional scatter-plot representation of a PCA and look at their evolution over time, some interesting connections can be revealed. Both conditions that follow random recommendations end quite soon in a consensual space of the originally dominant, unaligned community. Interestingly, this is not true for the personalized conditions. Here the two marginal communities spread out with the difference between \mathcal{L}_1 and $\mathcal{L}_{1|2}$, however, being that $\mathcal{L}_1 \times \text{Personalized}$ becomes almost egalitarian, as can be seen from the strong compression of the representations. $\mathcal{L}_{1|2} \times \text{Personalized}$, on the other hand, forms two clearly separated positions. This is underlined by examining the example graph representations (Fig. 3(b) to 3(d)). In episode 50, although the separation of communities for $\mathcal{L}_1 \times \text{Personalized}$ is still present, the focus of the debate is clearly at the center. In the case of $\mathcal{L}_{1|2} \times \text{Personalized}$, however, condensations are found in the focal points of the respective communities.

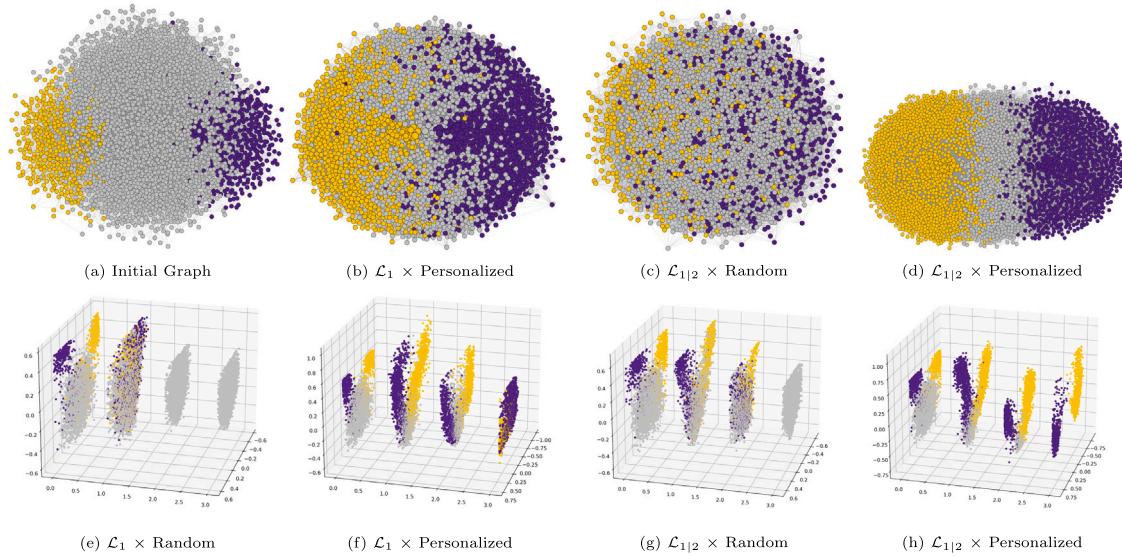


Fig. 3. Exemplary visualizations of graph topologies and latent space polarization for the divisive scenario. The graphs shown (Fig. 3(b) to 3(d)) are each taken from one simulation run at epoch 50. Initially (Fig. 3(a)), two small communities (depicted in violet and yellow) are integrated to some degree into a larger community (depicted in gray) while forming at opposite edges of the social graph. In condition $\mathcal{L}_1 \times \text{Personalized}$, both peripheral communities spread out, while in $\mathcal{L}_2 \times \text{Random}$, they become further integrated into the gray community, thus loosening internal connectivity. Concerning condition $\mathcal{L}_2 \times \text{Personalized}$, we observe pronounced centrifugal tendencies with condensed structures forming separately in the focal points of the respective communities. Below (Fig. 3(e) to 3(h)), polarization in latent space is depicted. Due to the presence of not only two, but three communities, we depict latent user distributions in form of a three-dimensional scatter plot. Concretely, we show the migration of user distributions by plotting a two-dimensional PCA over time at epochs 1, 25, 50, and 100. Clearly, only condition $\mathcal{L}_2 \times \text{Personalized}$ shows bipolarization while the remaining converge to consensus.

4.3. Discussion

Pitting the two scenarios against each other, our results clearly indicate that the initial structure of the debate space as well as personal ideological affiliations are essential for setting polarizing effects in motion (see also Appendix D).

The partisan scenario is characterized by a clear demarcation between two opposing communities from the outset. The communication graph is already somewhat delineated, with limited cross-cutting interactions. An illustrative real-world example is the political landscape of the United States, particularly issues such as “gun control” or “abortion rights (for a detailed alignment of the partisan scenario with real-world data, see Appendix C.1). Here, individuals typically have strong pre-existing beliefs and affiliations that align with either the Democratic or Republican parties. The debate is not about convincing the undecided, but about reinforcing existing beliefs and rallying one’s base. Notably, while the partisan alignment in this scenario is pronounced, we deliberately chose a relatively high proportion of cross-cutting connections. This decision was made to demonstrate that our model can operate at this limited degree of separation to model progressive divergence, approaching values found in real-world debates.

The results on the partisan scenario show that if the social graph is already ideologically structured at the start, the social polarization driver in itself leads to a fragmentation of the debate. Even if the recommendation space is very large, as in the case of $\mathcal{L}_2 \times \text{Random}$, centrifugal effects can occur in that messages from perceived outsiders are less likely to be spread. Thus, if personal identification with a social group is strong and sufficiently relevant to communication, oppositional fronts will sooner or later form, regardless of how wide or narrow individual information channels are. This corresponds to the perception of debates in the real world, which are sometimes conducted along ideological predispositions, such as party affiliation.

On the other hand, even though personalized message distribution densifies certain channels, the initial separation appears to be transient when individual demarcation alone is considered ($\mathcal{L}_1 \times \text{Personalized}$). This suggests that technological factors alone may not account for the bifurcation of discussion spaces. In cases where homophilic behavior intersects with personalization technologies, the outcome is not a

marked creation of epistemic gaps as predicted by the echo chamber hypothesis. Instead, it leads to a postponement in consensus building. Our findings corroborate studies arguing that digital platforms can potentially bridge preference gaps over time, despite users’ individual content selection [8,14,15].

Comprehending social polarization more holistically consequently necessitates considering not only the technological apparatus but also the social demarcation aspect. However, the ideological leanings of communication participants may not always be as transparent as in the first scenario. Given the opacity of personal filter bubbles and individual decision criteria for communicative behavior, determining the reasons for the greater divergence in some debates remains challenging. The uncertainty surrounding the roots of these divergences – whether they originate from human communication, dissemination technology, or a blend of both – highlights the complexity in identifying the causal mechanisms of social polarization.

In our second scenario, the unaligned case, the initial state of the debate within a particular network structure is characterized by ambiguity and a lack of clear ideological sorting. Contrary to the partisan scenario, where divisions are apparent from the outset, the unaligned scenario starts with a more consensual or neutral communication graph. A significant portion of the population initially share a view or have a neutral stance, while two smaller, more ideologically distinct communities exist at the fringes. Over time, as more information becomes available and narratives solidify, these peripheral communities can gain influence, progressively shaping the debate until they eventually become the dominant voices, leading to a polarized mainstream. Real-world examples reflecting this scenario include the initially broad consensus in the United States on supporting Ukraine in its war against Russia, which has since seen more divisive opinions emerge as the situation evolved. Certain aspects of the COVID-19 debates on social media can also be seen through this lens. Initially, there was a lot of uncertainty and ambiguity, but over time, as more information became available and narratives solidified, clear divisions emerged (for a detailed alignment of the unaligned scenario with real-world data, see Appendix C.2). Hence, such topological preconditions are particularly to be expected in case of novel or multifaceted subjects,

where initial discussions may be wide-ranging and the formation of personal opinions gradual.

A further complicating factor is that ideological sorting need not necessarily be clearly readable from a particular network structure. Particularly when novel topics of special importance or complexity are being negotiated, it can be assumed that, first, the debate is likely to be conducted very broadly, perhaps around a center of moderate opinion leaders, and, second, the formation of personal opinions does oftentimes not occur abruptly.

In our model, even when highly condensed communities emerge at the extreme edges, under a broad recommendation spectrum they fail to maintain their internal stability and become increasingly marginalized and ultimately incorporated ($\mathcal{L}_2 \times \text{Random}$). Interestingly, in contrast, the personalization of recommendations results in the two marginal communities prevailing rather than the original dominant one. This is especially remarkable for condition $\mathcal{L}_1 \times \text{Personalized}$, since participants do not exhibit explicit group-referential behavior in this case. According to our interpretation, the main reason responsible for this is the initially very strong networking within the marginalized communities (again compare [Appendix D](#)). This ensures that personal recommendations of members mostly originate from within their own community, thus provoking ongoing internal stability. The result, however, is by no means a fragmented, but rather a plural consensual space of opinions.

Social polarization occurs only if, in addition to personalized recommendations, ideological demarcation is achieved through communication. Accordingly, a key finding from the present study is that the causal relationships of social polarization cannot be grasped in their entirety if one merely considers social and technological parameters separately. Rather, we conclude that progressive demarcation of ideological groups is the result of the complex interdependence between both aspects.

As a consequence, accounting for group identity seems to be an important factor in modeling social polarization in a multidimensional setting adequately. It is widely accepted that the cultivation of a personal identity through differentiation from other social groupings is an essential feature of social organization. Thus, people perceive one and the same action very differently depending on whether they identify an actor as a member of their own or of a foreign group [53]. Consequently, actions by members of one's own group are generally evaluated more favorably than those observed by outsiders. This discrepancy even manifests itself when people are randomly assigned to a group for no objective reason (cf. minimal group paradigm [54]).

According to Laclau and Mouffe, the antagonistic aspect underlying the identity formation of groups is a constitutive feature of the political process [76]. They state that a harmonious society is ultimately impossible to achieve since the systematic distinction between friend and foe determines to an existential degree how a society functions. Accordingly, the negotiation of ideological differences is in the last instance an inherent dimension of any social practice. This structural antagonism, thereby, extends far beyond the sphere of politics in the narrow sense; rather, it is the fundamental difference of any ideologically shaped social form. According to Mouffe, the irrevocability of the antagonism of political actors is the central deficit of liberal theories of democracy, which assume that political antagonisms will disappear through greater individualization and rational discourse [77].

The disruptive impact of ideologically shaped friend/foe differentiation shows itself in the formation of deficient patterns of cognition. From an epistemological perspective, ideology describes epistemically deficient, i.e. distorted or illusionary, thinking, which is formed by a reference to a specific system of ideas and values [56]. It is a cognitive phenomenon of need-based thinking that produces misconceptions influenced by interests, needs, and desires. Hence, an ideology can be regarded as both an individual quantity or a world of ideas shared by a group. Although ideologies are always formed in social frames of reference, individual people are the bearers of the resulting structures. This means that not only explicitly formulated major ideologies with far-reaching group integration are to be recognized as ideologies, but in

general social systems that are structurally coupled to the cognitive process of their members through the mediation of ideological fragments or signs.

In our model, this relationship is expressed by the fact that the probability of intra- and inter-group communication is moderated by the expectation that a receiver directs at the sender of a message. The prior identifies the latter on the basis of categorical schemata as a member of one's own or of a foreign group, provoking the psychological phenomena of group behavior discussed earlier. The resulting notion of boundary is thus an answer to the question of how, in the face of ongoing cross-border communication and constantly reproducing surpluses of meaning, an ideological group can nevertheless persist. It is true that ideological network formation generally appears to be a highly precarious process, since it is initially subject to arbitrary addressability and social networks have a broad distribution spectrum of communication at their disposal. Ideological networks, however, find support in themselves, precisely in the particularity of the linkage and in the interlocking of the unique social structures that apply to them.

At the same time, it must be ensured that technologically moderated communication channels permit and safeguard these processes of ideological network formation by creating the social redundancy necessary to stabilize intra-group communication. By social redundancy, we mean that the same people keep coming together over time in different functions, e.g. as authors, recipients, or disseminators of messages, to produce a certain surplus of shared ideological fragments. As platform algorithms operate as black boxes, however, it is almost impossible to determine to which degree channels of communication have been consensed or expanded. The purposeful transformation of information into redundancy, which can be understood as a central mechanism of *echo chambers*, is hence difficult to comprehend. The primary problem is not that digital corporations shy away from revealing the algorithmic foundation of their dissemination technology, but rather that social redundancy is anonymized rendering the control effect of recommendations opaque. For one thing, it is unclear to the creator of a message which individuals received their messages in the first place. For another, the receiver may not know which other (possibly contrastive) messages circulate outside their information horizon. In this sense, the generation of technological horizons of information as well as the scope of social redundancy describe the core problem of filter bubbles.

It is important to emphasize that there is a potential trade-off between the economic interests of platform companies and the liberal democratic aspiration to guarantee and ensure access to a balanced diet of information. The curation of the news feed as the central interface between people and platform relies on predictive algorithms that extract myriad parameters from the behavioral surpluses of platform users to calculate the personal relevance of thousands of eligible messages favoring those from people who have been interacted with in the past, messages that garnered lively interest from others, and messages that are similar to those that were of interest to a user at some point [78].

The goal of digital platforms is to create a closed loop that feeds off the needs and inclinations of its users, amplifying and then potentiating them. When we talk about needs in this context, we are primarily referring to expected values extracted from behavioral data, which do not necessarily overlap with a person's real needs. Thus, personalized content may suggest that needs are being met and increase the (emotional) involvement of a user in the short term, but it has been frequently contested whether positive effects also materialize in the long run [79–81]. With regard to the role of social platforms in the shaping of public opinion, we argue that algorithmic dissemination may in particular lead to a progressive condensation of information channels, which in turn runs counter to undistorted information gathering. In our view, personalized news feeds thereby reproduce (latent) ideological structures thus favoring the emergence of ideologically congruent networks.

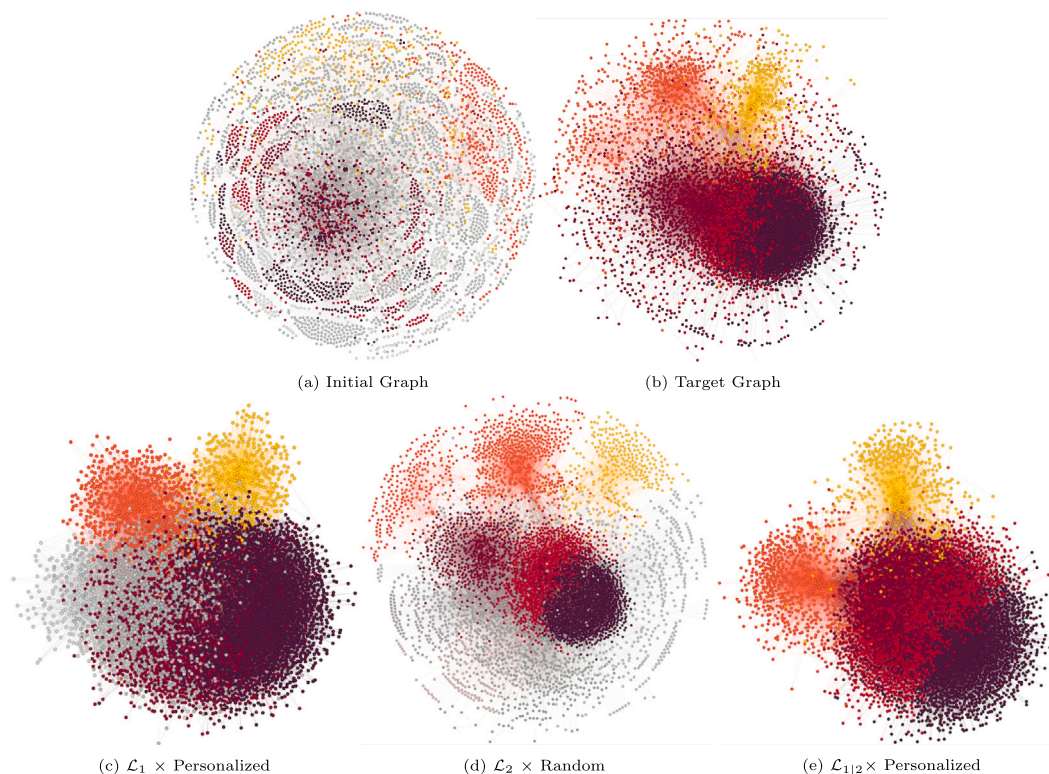


Fig. 4. Visualizations of graph topologies on a real-world dataset. The initial graph (Fig. 4(a)) consists of 20% of the earliest edges contained in the target graph (Fig. 4(b)). While the initial graph has very loose connections and thus a large number of communities is identified, the target graph is much more condensed and contains only 5 communities. We calculated Jaccard similarity to the target network's communities to assign colors to the communities in the initial as well as the scenario graphs. All three depicted scenarios (Fig. 4(c) to 4(e)) capture to some degree the properties of the target graph. For instance, they all identify the most clearly separated communities (in orange and yellow). However, at closer inspection $\mathcal{L}_1 \times$ Personalized shows too much and $\mathcal{L}_2 \times$ Random too little condensation. Although there are still obvious topological discrepancies to the target graph, we view $\mathcal{L}_{1/2} \times$ Personalized as achieving the closest approximation.

5. Conclusion

In studying the evolution of communication on an idealized social media platform using agent-based modeling, we conceptually follow a rich tradition of methods subsumed under the term opinion dynamics. In contrast to existing methods, however, we rely on the predictive power of modern machine learning methods and the semantics derived by them.

The findings presented here suggest that our method is able to detect and simulatively extrapolate community-related structures at an early stage of a debate. Our approach thereby directly operates on an existing graph structure, thus making comparisons to empirical graphs especially convenient. For one, deriving latent spaces through machine learning relieves us from the necessity to position participants in an explicit (one-dimensional) attitude space in order to capture social polarization. Thus, latent polarization patterns can be derived from multidimensional behavioral data, which can be poignantly visualized via dimension reduction procedures. From this, complex migration patterns between several communities can be analyzed over time (cf. Fig. 1(e) to 1(h) and 3(e) to 3(h)), which allows more meaningful conclusions about the progression of social polarization than the evaluation of the graph topology via corresponding metrics such as modularity or homophily. Moreover, the need to manually align simulation parameters with empirical data, such as the number of actors or the degree of compaction of the expected communities [cf. 33, 41, 44], is eliminated.

Consequently, our method provides a flexible tool to determine how, for instance, varying the recommendation technology might affect the evolution of network communication. The concept of conjunctive graph queries offers a great deal of leeway in the formulation of such hypotheses [65]. In the present work, only rather simple operations were applied in order to convey the basic ideas clearly. In principle,

however, far more complex queries may be posed, for example to design sophisticated recommendation algorithms. These can then be used to analyze how, for instance, drawing messages from either the centers of communities or from the boundary region between them may affect the development of social polarization [48]. Apart from that, the framework is sufficiently flexible to integrate further edge types and edge type combinations as well as geometric operators without much effort. A useful addition in this sense would possibly be the consideration of other forms of social interaction such as liking or replying. Perspectively, the integration of argumentation graphs may be conceivable.

Beyond, there is still much room for improvement to the proposed method. For example, to properly apply our approach to real-world data, we need to deal with identifying structural or content predictors that allow us to make assumptions about the conditions under which a certain community will exhibit appropriate ideological demarcation. Overall, our method is very dependent on the type of community detection method used. Further research should be carried out in this regard in the future. In addition, our model lacks some key features that make it difficult to apply it in real-world situations. These include, first, that it is currently not possible for new actors or messages to enter the discourse. Second, it seems reasonable to us to derive latent representations across different topics. Ideological group effects often show up consistently over various contexts, such that specific communication patterns can likely be transferred and generalized to yield more robust predictions. Third, we are aware that integrating computational linguistic methods will be an important step for future work since taking into account the actual message content is essential to model ideologically shaped behavior.

Finally, and most importantly, based on synthetic data alone, we cannot determine the extent to which our findings generalize to real-world scenarios. Empirical network graphs are usually characterized by

a much more complex structure, consisting, at the microscopic level, of more fine-grained local communication and, at the mesoscopic level, of a larger set of communities that are interconnected to varying degrees. It is rarely the case in reality that a debate is so clearly divided into two (or three) viewpoints as in the present synthetic studies. Therefore, in the future, we will extend the insights we have gained in this work and thoroughly apply them to real-world empirical data. Again, see Fig. 4 for initial promising results on a social graph extracted via the Twitter-API and Appendix F for further descriptions.

CRedit authorship contribution statement

Tim Donkers: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Data curation, Writing. **Jürgen Ziegler:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Latent space model

In this section, we present the graph-based embedding model that we have applied to an opinion dynamics framework in the main text. We (i) introduce a general way to formalize graphs, to (ii) be able to express social network graphs accordingly. Then we (iii) show how to fit such graphs into a differentiable framework that lets us represent graph nodes and edges in a latent semantic embedding space.

A.1. Preliminaries

We define $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as a graph consisting of nodes $v \in \mathcal{V}$ and edges $e \in \mathcal{E}$. Edges are defined as binary predicates $e = \tau(v_1, v_2)$ with $v_1, v_2 \in \mathcal{V}$ and a relation $\tau \in \mathcal{R}$ of a specific type such that $\tau : \mathcal{V} \times \mathcal{V} \rightarrow \{0, 1\}$ describes an edge relation. Hence, in the following we assume heterogeneous graphs with multiple different types of nodes and edges. When referring generally to nodes, we use the letter v with varying subscripts; however later on, where type information is salient, we will use distinct letters to denote nodes of different types, e.g. m for a message node in a communication graph. Finally, we use lower-script, e.g. m , for actual graph nodes and upper-case, e.g. M , for variables whose domain is the set of graph nodes.

A.2. Social networks

A heterogeneous communication graph \mathcal{G}_c is comprised of a set of user nodes $\mathcal{V}_u \subset \mathcal{V}$ and message nodes $\mathcal{V}_m \subset \mathcal{V}$. Since the primary focus of this work is the simulation of connecting communication (such as liking, replying or reposting), we define communicative edges from the perspective of the recipient of a message as $\tau_r(u, m), u \in \mathcal{V}_u, m \in \mathcal{V}_m$. The set of edges is accordingly $\mathcal{E}_r = \tau_r(u, m) | \{ \tau_r(u, m) = 1 \wedge m \in \mathcal{V}_m \wedge u \in \mathcal{V}_u \wedge \tau_r \in \mathcal{R} \}$. Informally, we can say ‘User u reacted to message m ’. For the act of sending a message, we inversely choose $m \in \mathcal{V}_m$ as the head component and author user $u \in \mathcal{V}_u$ as the tail such that the set of edges is defined as $\mathcal{E}_s = \tau_s(m, u) | \{ \tau_s(m, u) = 1 \wedge m \in \mathcal{V}_m \wedge u \in \mathcal{V}_u \wedge \tau_s \in \mathcal{R} \}$ with $\tau_s(m, u)$ being the edge relation. Informally, this edge type describes the relation ‘Message m was sent by user u ’. Put together, the unmodified communication graph up to this point is defined as $\mathcal{G}_c = (\mathcal{V}_u \cup \mathcal{V}_m, \mathcal{E}_r \cup \mathcal{E}_s)$.

A.3. Conjunctive graph queries

Having derived the basic nodes and edges, we now want to exploit the topological information to infer latent relationships and make predictions about the future evolution of the graph. Specifically, our intuition is to derive a predictive model based on statistical correlations between nodes as well as the different types of edges. In the simplest case, we hide the tail component of an existing edge $\tau(v_1, v_2)$ and try to predict it based on head and edge type. However, we also aim at making predictions about more sophisticated queries where several edges can be concatenated.

From the perspective of graph logic, we are interested in providing answers to conjunctive graph queries. We, hence, follow [65] to write queries $q \in Q(\mathcal{G})$ as:

$$q = V_\tau. \exists V_1, \dots, V_n : e_1 \wedge e_2 \wedge \dots \wedge e_m, \quad (A.1)$$

where $e_i = \tau(v_j, V_k), V_k \in \{V_\tau, V_1, \dots, V_n\}, v_j \in \mathcal{V}, \tau \in \mathcal{R}$
or $e_i = \tau(V_j, V_k), V_j, V_k \in \{V_\tau, V_1, \dots, V_n\}, j \neq k, \tau \in \mathcal{R}$,

where V_τ denotes the target variable of the query, i. e. the node(s) that we want the query to return, while V_1, \dots, V_n are existentially quantified bound variable nodes. The edges e_i can involve these variables as well as constant anchor nodes.

Given the graph queries defined in Eq. (A.1) and due to our definition of edges as binary predicates, we are faced with the restriction that we have only a single free variable as the sink node. For social graphs, however, it is mostly true that multiple true responses to a single query exist; for example, because users typically respond to more than one message over time. Furthermore, we are particularly interested to predict future and thereby previously unobserved edges. Therefore, with regard to a query $q \in Q(\mathcal{G})$ we need to distinguish between the observed denotation set $\llbracket q \rrbracket^o$, which contains all nodes that exactly satisfy q , and the unobserved (hidden) denotation set $\llbracket q \rrbracket^h$ that we want to predict.

Summarized, our goal is to exploit a given graph topology by constructing queries and coupling them with observed denotation sets yielding question-answer pairs $(q, v^*), v^* \in \llbracket q \rrbracket^o$. Feeding this data into a prediction model consequently helps us uncover generalizable structures which, in turn, can be utilized to make predictions about missing edges.

A.4. Knowledge graph embeddings

The model of graph-based queries described above is essentially based on the assumption that semantic relationships between graph entities can be captured in the form of a conjunctive linkage of edges. However, since edges of the same type can exist between a range of different nodes, this implies in particular a generalizability of these semantics. Our hypothesis is now that this semantic generality can be expressed and operationalized in terms of latent information spaces.

Pioneering work on the quantitative representation of semantic contexts has already demonstrated that appropriately trained latent spaces are organized in such a way that they position semantically similar entities in close proximity to each other [cf.64]. In particular, it follows that the distances between entities incorporate latent semantics as well. We explicate this connection in the following by representing queries as logical geometric operators that are optimized jointly with the vector representations of graph nodes.

For example, we will show that, starting from an embedded user node, we can apply geometric operations to identify a region in information space where message nodes are in close proximity to each other that this user has a high probability of propagating.

Formally, we map every conjunctive input query q to a d -dimensional embedding vector $\mathbf{q} \in \mathbb{R}^d$ by applying differential operators to the set of input anchor nodes $\{v_1, \dots, v_k\}$ extracted from the DAG of q . These operators (defined below) are trained alongside the embeddings of graph nodes $\mathbf{v} \in \mathbb{R}^d, \forall v \in \mathcal{V}$.

Distance operator. After applying the relevant geometric operations the resulting embedding \mathbf{q} can be used to predict the likelihood that any particular node $v \in \mathcal{V}$ satisfies query q . In particular, our learning process is designed such that the likelihood that $v \in \llbracket q \rrbracket$ is expressed by a distance function $\delta : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. Concretely, we define the distance between query and a node embedding as:

$$\delta(\mathbf{q}, \mathbf{v}) = \frac{1}{2} \frac{\text{Var}(\mathbf{q} - \mathbf{v})}{\text{Var}(\mathbf{q}) + \text{Var}(\mathbf{v})}, \quad (\text{A.2})$$

where Var is the variance of the input vector. Our training objective is, thus, to generate a query embedding that implicitly represents its denotation set $\llbracket q \rrbracket = \llbracket q \rrbracket^o \cup \llbracket q \rrbracket^h$ so that $\delta(\mathbf{q}, \mathbf{v}) = 0, \forall v \in \llbracket q \rrbracket$ and $\delta(\mathbf{q}, \mathbf{v}) = 1, \forall v \notin \llbracket q \rrbracket$ is approximated. To achieve this, the parameters of the geometric operators have to be updated subject to the generalizable semantic commonalities found in the graph topology.

Projection operator. The geometric operator that we utilize frequently throughout this paper is geometric projection. The projection operator $\mathcal{P} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is applied to a query embedding \mathbf{q} subject to an edge type τ . Hence, \mathcal{P} outputs a new query embedding $\mathbf{q}' = \mathcal{P}(\mathbf{q}, \tau)$ with the denotation set $\llbracket q' \rrbracket = \bigcup_{u \in \llbracket q \rrbracket} \mathcal{N}(u, \tau)$ with $\mathcal{N}(u, \tau)$ being the set of nodes connected to u by edges of type τ . Since the actual implementation of \mathcal{P} can vary a lot and although there exist quite sophisticated approaches (see [82] for a comprehensive overview), we want to stay within the picture of translating embeddings in a semantic space to convey our concept as clearly as possible. Hence, we define \mathcal{P} according to [83] as an addition of a query vector \mathbf{q} and an edge-type specific translation embedding $\mathbf{r}_\tau \in \mathbb{R}^d$:

$$\mathcal{P}(\mathbf{q}, \tau) = \mathbf{q} + \mathbf{r}_\tau \quad (\text{A.3})$$

Please refer to [48,65] for further information on and implementations of geometric operators in the context conjunctive graph queries.

Summarized, we (i) represent any target query q in form of its DAG, (ii) start with the embeddings $\mathbf{v}_1, \dots, \mathbf{v}_n$ of the anchor nodes, then (iii) apply geometric operators \mathcal{P} to these embeddings to obtain \mathbf{q} at a position in latent space where we assume nodes that satisfy q , and finally, during our training process, (iv) calculate the distance of \mathbf{q} towards the embedding of a node $v \in \llbracket q_i \rrbracket^o$ observed in the data. As a corresponding objective function for the learning process, we utilize a max-margin loss with negative sampling:

$$L(q) = -\log(y - \delta(\mathbf{q}, \mathbf{v})) - \sum_{i=1}^k \log \sigma(\delta(\mathbf{q}, \mathbf{v}'_i) - y), \quad (\text{A.4})$$

where $y \in [0, 1]$ depicts a fixed scalar margin, σ is the logistic function, and $k \in \mathbb{N}$ is the number of negative samples.

Appendix B. Synthetic graph generation

Since the degree of network polarization is determined based on the relations that exist between users, we start by producing synthetic user graphs \mathcal{G}_u . In a subsequent step, we then derive a communication graph \mathcal{G}_c based on the resulting user connections. We first assume that a user graph consists of more or less strongly connected communities. Accordingly, we first define a number of communities of a certain size with a variable degree of internal and external connectivity. Furthermore, empirical findings of real social networks suggest that although egalitarian exchange may take place on social platforms in principle, clear hierarchical structures can still be identified as a rule [e.g.84]. Consequently, a significant portion of network communication is carried out via a rather small set of users, whom we call gatekeepers. We distinguish them dichotomously from the set of regular group members by assigning them each a higher number of communication units directed at them.

Formally, given a community $c \in \mathcal{V}_c$, we write its set of regular member nodes as $\mathcal{V}_u^m[c] \subset \mathcal{V}_u$ and its gatekeepers as $\mathcal{V}_u^g[c] \subset \mathcal{V}_u$ such that $\mathcal{V}_u[c] = \mathcal{V}_u^m[c] \cup \mathcal{V}_u^g[c]$ is the complete set of nodes associated with community c .

For a given user $u \in \mathcal{V}_u[c]$, we define the number of users connected to them in the user graph $|\mathcal{N}(u, \tau_u)|$ to be drawn from a Poisson distribution with the expected value chosen subject to u being a regular member or a gatekeeper:

$$|\mathcal{N}(u, \tau_u)| \sim \begin{cases} \text{Poisson}(\psi_c^m) & \text{if } u \in \mathcal{V}_u^m[c] \\ \text{Poisson}(\psi_c^g) & \text{if } u \in \mathcal{V}_u^g[c], \end{cases} \quad (\text{B.1})$$

where ψ_c^m and ψ_c^g is the expected number of connections for members and gatekeepers respectively.

Whether an edge is to be produced within or between communities is determined on the basis of an identically and independently distributed categorical distribution with replacement:

$$C_1, \dots, C_{|\mathcal{N}(u, \tau_u)|} \sim \text{IID Cat}(\mathbf{p}_c) \quad (\text{B.2})$$

$$\mathbf{p}_c := (p_1^c, \dots, p_{|\mathcal{V}_c|}^c),$$

where p_k^c is the probability of choosing community c_k given that user u is a member of community c .

In the next step, without restriction of generality and starting from the identified community variable C_j , an index is sampled from the set of member nodes $\mathcal{V}_u[C_j]$ using a discrete power law distribution:

$$I_{C_j}(U_j) \sim \text{DPL}(a, |\mathcal{V}_u[C_k]|), \quad (\text{B.3})$$

where $I_{C_j} : \mathcal{V}_u[C_j] \rightarrow \{1, \dots, |\mathcal{V}_u[C_j]|\}$ is a bijective mapping between indices and member nodes of C_j such that U_j is the resulting user node when applying $I_{C_j}^{-1}$ to the sampled index. Furthermore, $a \in \mathbb{R}$ is some constant center skew (which we set to 1 for all scenarios to achieve a pronounced hierarchical structure), and $|\mathcal{V}_u[C_k]|$ depicts the number of different indices that can be returned by the power law function for a given community. Informally, Eq. (B.3) accounts for the fact that users differ in how frequently they engage in connecting communication.

From the resulting randomly generated user graph, we, in the next step, derive the communication graph \mathcal{G}_c . For a given user u , we define the number of messages sent $|\mathcal{N}(u, \tau_s)|$ to be drawn from a Poisson distribution:

$$|\mathcal{N}(u, \tau_s)| \sim \text{Poisson}(\psi), \quad (\text{B.4})$$

where $\psi \in \mathbb{N}$ is the expected number of messages sent. Next, for each message m to be sent, we sample a number of n users from the set of users connected to u to establish edges $\tau_r(U, m)$ where n depicts a lower-bound, that is the minimum number of reactions towards a message. Please note that we choose U such that, at least, one edges is established for each connected user to make sure that there is no mismatch between communication graph and user graph. Additionally, since our training procedure is sensitive to data sparsity, we define an expected minimum number of reactions by each user. If the generated graph is too sparse, we extend it by densifying it accordingly.

Appendix C. Real-world alignments of synthetic scenarios

This section attempts to match our proposed scenarios, represented by specific social network graph topologies – the partisan and the unaligned – with tangible real-world data. The primary goal is to validate these scenarios, emphasizing their ability to accurately represent the initial or very early stages of real-world debates. By juxtaposing synthetic datasets with real-world instances, we aim to illuminate the nuanced patterns of polarization that emerge in digital discussions.

C.1. Partisan scenario

The partisan scenario is emblematic of situations in which two or more opposing communities are to some extent demarcated from the outset. In such settings, the debate is less about swaying the undecided and more about reinforcing existing beliefs. A quintessential real-world manifestation of this scenario can be observed in the political landscape of the United States, particularly in debates over contentious issues such as ‘‘climate change’’ or ‘‘abortion rights’’.

The phenomenon of social polarization is becoming increasingly relevant and pronounced. According to the Pew Research Center, some 62% of Republicans and 54% of Democrats have a very unfavorable view of the opposing party [85]. While these highly negative views have remained consistent in recent years, they have seen a significant uptick from just five years ago and are notably higher than a few decades ago. In 1994, less than a quarter of either party expressed such strongly unfavorable views of the other party.

Further evidence from Smidt [86] underscores this trend. The research shows a significant decline in the rate at which Americans vote for different parties across successive presidential elections. The clarity of party differences, exacerbated by polarization, has reduced indecision and ambivalence, leading to increased reliability in presidential voting. Even independent and less engaged voters now perceive candidate differences as clearly as partisan, engaged voters in past elections. This clarity of party differences has led to declining rates of ambivalence, indecision, and swing voters. In particular, pure independents have become as reliable in their party support as strong partisans of earlier eras.

This trend towards polarization, however, is not limited to the United States. When we turn our attention to Europe, we find similar patterns. Polarization in Europe is not a recent phenomenon. Post-World War II data show that polarization, especially electoral support for radical political entities, has been escalating since the 1960s [87]. However, it was in the 1990s, with the transformation of European party systems and the rise of (populist) radical parties, that votes for anti-political establishment parties rose to unprecedented heights.

In recent decades, the vote share for anti-political establishment parties, be they populist, anti-system, protest, radical, or extreme, has grown exponentially. This increase in polarization is evident in both Western and East-Central European countries. In Western Europe, for example, there has been a significant increase in votes for these parties, especially in countries facing challenges such as immigration (e.g., Austria, France, Germany, the Netherlands) or economic crises (e.g., Spain, Italy, Greece, Cyprus) [87]. On average, polarization has increased by more than five percentage points over the past decade. Compared to less polarized periods, such as the 1960s, recent years have seen polarization nearly triple.

In Asia, India's political landscape offers a vivid example of polarization. The rise of the Bharatiya Janata Party (BJP) and its Hindu nationalist agenda has deepened divisions between religious communities, particularly Hindus and Muslims [88]. Similarly, in Africa, countries such as Kenya have witnessed increased polarization, particularly during election periods. Kenya's 2017 general election, for example, saw deep divisions between supporters of the two main political factions, leading to widespread protests and violence [89]. Such instances underscore that the phenomenon of polarization is not limited to the West, but is a global trend.

The growing divergence observed not only in the United States and Europe, but also in regions such as Asia and Africa, underscores the broader relevance of social polarization. This phenomenon is not limited to abstract political ideologies, but manifests itself concretely in debates on pressing global issues across continents. One such issue that has become emblematic of this divide, transcending national and regional boundaries, is climate change.

In the United States, climate change has moved from a scientific concern to a major political issue. The issue, which ideally should be addressed with a unified front given its global implications, has instead become a battleground for political ideologies. Currently, 54% of U.S. adults perceive climate change as a major threat to the nation's well-being [90]. While this number has declined slightly since 2020, it remains significantly higher than it was in the early 2010s. The partisan divide is evident when these numbers are broken down: 78% of Democrats see climate change as a major threat, a significant increase from 58% a decade ago. In stark contrast, only 23% of Republicans

share this view, a number that has remained stagnant over the past decade.

The divide deepens when it comes to the causes of climate change. A large majority of liberal Democrats (79%) attribute the Earth's warming primarily to human activity [91]. By contrast, only 15% of conservative Republicans share this view, with a significant share either denying the evidence of global warming (36%) or attributing it to natural causes (48%). This polarization is not a new phenomenon. Already in 2015, there was a 41-point gap between Democrats and Republicans on the human contribution to climate change.

While longstanding issues like climate change have been the subject of partisan divisions for years, new issues can also quickly become points of contention. The COVID-19 pandemic is a prime example of this phenomenon. As the world grappled with an unprecedented public health crisis, the response to the pandemic almost instantly became a point of political contention in many countries, with Brazil being a particularly striking example [92].

Then-President Jair Bolsonaro's approach to the pandemic was marked by a dismissive attitude towards the seriousness of the virus. He actively obstructed Brazil's federal agencies, urged mayors and state governors to rescind stay-at-home orders, and consistently resisted calls for social distancing. Bolsonaro's supporters echoed his sentiments, sharing his social media posts, defying stay-at-home orders, and minimizing the health risks of the crisis. In stark contrast, the opposition, the media, and most health professionals criticized the president for his polarizing messages that failed to adequately address the health crisis.

This polarized response to the pandemic was not unique to Brazil. In the United States, the partisan divide over the perceived threat of COVID-19 is evident. A recent Pew Research Center survey found that 85% of Democrats and Democratic-leaning individuals view the coronavirus as a major threat to the nation's well-being [93]. In contrast, only 46% of Republicans and Republican leaners share this view, with 45% viewing it as a minor threat. This gap has remained consistent even as the number of COVID-19 cases has skyrocketed in various states.

Cross-country comparisons further illustrate the extent of this polarization. A study comparing the United States, Canada, and the United Kingdom found that political conservatism was associated with misperceptions about COVID-19 in all three countries [94]. However, this association was significantly stronger in the US than in the UK. The study also found that polarization appears to have increased significantly between March and December 2020, particularly in the US. Nevertheless, polarizing tendencies were evident from the outset.

As another example, in India, the announcement and subsequent enactment of the National Register of Citizens (NRC) and the Citizenship Amendment Act (CAA) quickly polarized the nation [88]. Social media platforms, particularly Facebook and Twitter, became hotbeds of debate, mobilization, and protest organizing. The rapid rise of online groups and pages dedicated to the issue, coupled with significant spikes in membership following key events, highlighted the immediacy with which contemporary issues can become divisive. Much like the case of COVID-19, the NRC-CAA protests underscore a global trend of rapidly polarizing issues, with digital platforms playing a central role in shaping public opinion and driving divisions.

Brexit, the United Kingdom's monumental decision to leave the European Union, serves as another compelling example of how quickly issues can become polarized [95,96]. Despite the ebb and flow of political dominance over the years, the Brexit referendum was already entrenched as a deeply divisive issue from the start in 2013. Of particular note is the stability of the undecided voter demographic. In the years leading up to the 2016 referendum, the proportion of those who had not yet made up their minds remained consistently below 20%.³ This suggests that while the Brexit debate was to some extent characterized by an environment of uncertainty, the issue drew lines

³ Data available at <https://www.whatukthinks.org/eu/>.

of division from the outset. This early polarization underscores the inherently contentious nature of Brexit, with deeply held opinions and beliefs shaping the discourse long before the actual vote.

The speed with which debates polarize on online platforms is strikingly evident when examining the discourse surrounding contentious issues. One example is the aftermath of the 2009 shooting of Dr. George Tiller, a prominent figure in the U.S. abortion debate [97]. Analyzing Twitter data from the first 24 h after the incident, the researchers found that the platform quickly became a breeding ground for both pro-life and pro-choice discussions. The majority of users who participated in the conversation had already aligned themselves with either the pro-life or pro-choice camp, with very few remaining neutral or undecided. This immediate alignment underscores the power of pre-existing beliefs and affiliations in shaping online discourse.

Interestingly, pro-choice believers were nearly three times more likely to respond to other pro-choice believers, and a similar pattern was observed among pro-life believers. This trend of like-minded individuals engaging with one another highlights the echo chamber effect, whereby individuals are more likely to interact with those who share their views, thereby reinforcing their beliefs.

Moreover, the study highlighted that even in the face of a tragic event, the debate did not revolve around the incident itself, but was deeply rooted in the broader pro-life/pro-choice divide. The immediacy of the polarization, especially on an issue historically known for its partisan divide, demonstrates how online platforms can both reflect and amplify real-world divisions. Such findings underscore the importance of understanding the dynamics of online debates, as they offer a window into societal divisions and the speed with which they can manifest in the digital realm.

C.2. Unaligned scenario

While the partisan scenario offers a clear delineation between opposing communities from the outset, the landscape of online debate is not always so straightforward. The unaligned scenario presents a more nuanced picture of online discourse, where the initial state of a debate may not be as clearly delineated. This scenario is characterized by a more egalitarian communication graph, where a significant portion of the population remains undecided and two communities exist at the edges. Over time, these fringe communities may increasingly influence the debate and eventually become the divided mainstream.

We argue that several factors may account for the potentially more ambiguous initial state of online debates:

Traditionally consensus topics. Some issues may have historically enjoyed broad consensus, or at least been viewed as neutral. However, as social dynamics change and new information emerges, these issues can become more polarized. For example, early discussions about climate change were more open, with many people undecided. As the debate progressed and became more heated, clear divisions emerged.

Ambiguity in new topics. New issues can introduce ambiguity, especially when it is unclear how different ideological groups might interpret them. The early stages of debates, such as the Russian War or the early days of the COVID-19 pandemic, were characterized by uncertainty. As more information became available and narratives solidified, clear divisions emerged.

Nature of the initial network graph. The structure of the initial communication graph itself may not have clearly delineated positions. Instead, it may reflect a more fluid and dynamic state of discourse in which positions and affiliations are still in flux.

In essence, while the partisan scenario provides a snapshot of debates that are polarized from the start, the unaligned scenario underscores the evolving nature of online discourse. It highlights how debates can begin in a more open and undecided state, only to become polarized as they progress. This dynamic nature of online debates

underscores the importance of understanding not only the end state, but also the path of polarization. Below we explore these factors in more detail.

The evolution of public opinion on traditionally consensual issues offers a compelling insight into the dynamics of polarization. Consider the example of climate change. Although we have already introduced climate change as a broadly partisan issue, historical data show that concerns about the detrimental effects of global warming were once shared across party lines. In the 1980s, as climate change began to gain traction as an issue of public concern, there was a noticeable shift in public perceptions. By 1989, 75% of U.S. respondents considered climate change a serious problem, up from 43% in 1982 [98]. This evolution is further underscored by the fact that the proportion of respondents who were uncertain about the issue fell from 37% at the beginning of the decade to just 11% in 1989, suggesting that the middle ground between the two positions had effectively disappeared. This consensus has since eroded, however, and the issue is now a clear point of division.

Similarly, the perception of immigrants as a threat to the U.S. was once a relatively divisive issue. Until 2002, Americans of all political affiliations were largely undecided on the issue, with 58% of Democrats and 56% of Republicans expressing this sentiment [99]. By 2019, however, the issue had morphed into a deeply partisan one, with only 19% of Democrats and a staggering 78% of Republicans viewing immigrants as a threat. This shift has been particularly pronounced during the Trump years, with the gap between supporters of the two parties widening markedly. For example, in 1998, the difference in views on whether immigrants posed a significant threat was only 2 percentage points. By 2020, however, the gap had widened to 48 percentage points.

Moreover, the U.S. Supreme Court, a revered institution, is not immune to these polarizing trends. Currently, less than half of Americans (44%) express a favorable opinion of the Court [100]. This decline in favorability is particularly pronounced among Democrats, with just 24% viewing the Supreme Court favorably, the lowest level in more than three decades. In stark contrast, 68% of Republicans hold a favorable view of the Court, underscoring the widening gap between the two parties.

These dynamics of polarization, as we have explored, often manifest themselves over extended periods of time, reflecting deep-seated societal shifts and evolving narratives. It is important to recognize, however, that the temporal scope of many studies of polarization in online social networks is much narrower. The trends and patterns we have highlighted, while spanning years or even decades, can also be observed in much more compressed time frames. This observation is crucial because it underscores the speed with which public opinion can polarize, even in the digital age.

A case in point is the recent Russian war on Ukraine. In January 2022, there was a bipartisan consensus in the U.S. that Russia was not perceived as a significant threat, with 27% of Republicans and 26% of Democrats holding this view [101]. Fast forward a year to January 2023, and the landscape has shifted: 29% of Republicans still hold this view, but the proportion of Democrats has risen to 43%. This divergence becomes even more pronounced when we examine attitudes towards aid to Ukraine. In March 2022, only a small fraction of both Republicans (9%) and Democrats (5%) thought the U.S. was providing too much aid to Ukraine. By June 2023, this sentiment had grown dramatically among Republicans, with 44% expressing this view, compared to just 14% of Democrats.

Taken together, these examples underscore a broader trend: issues that were once consensual or neutral ground are becoming increasingly polarized. Whether these issues were traditionally consensual or emerged as new and ambiguous issues, the pattern is clear. Over time, as narratives solidify and political cues become more pronounced, clear divisions emerge. This phenomenon is not limited to the U.S., but is a

global trend that highlights the ever-evolving nature of public opinion and the profound impact of polarization.

While the rapid polarization of issues such as Russia's war on Ukraine underscores the volatile nature of online debates, it is important to examine the broader implications of such divisions. Not only can an issue quickly become polarized, but the effects of that polarization can reverberate and influence perceptions and opinions on a range of related issues. Again, the discourse surrounding COVID-19 in the United States serves as a prime example of this phenomenon.

For example, in March 2020, significant majorities from both major political parties expressed trust in scientific agencies such as the Centers for Disease Control and Prevention (CDC) for information related to the coronavirus. Specifically, 74% of Republicans and 84% of Democrats held this view [102]. However, just a few months later, in July 2020, this confidence diverged dramatically. Only 25% of Republicans continued to trust these institutions, while trust among Democrats jumped to 89%.

This shift not only demonstrates the speed with which public opinion can change, but also underscores the interconnectedness of issues in public discourse. The polarization of one issue, such as COVID-19, can have cascading effects, influencing and polarizing opinions on other related issues. Such observations are crucial for researchers and policy makers alike. They underscore the need for a nuanced approach to studying online debates, as a broader issue and its specific facets can offer different insights into the evolving landscape of public opinion and the intricate web of connections that shape it.

Nigeria's political landscape offers another compelling illustration of the cascading effects of polarization. The election of President Muhammadu Buhari in 2015 set the stage for a pronounced divide between the Muslim and Christian populations. Initially, Buhari enjoyed widespread approval, with 94% of Muslims and 58% of Christians expressing support [103]. By 2018, however, this consensus had eroded dramatically. While 73% of Muslims still approved of Buhari, only 26% of Christians did - a stark contrast to the earlier figures.

But the impact of Buhari's election went beyond his personal approval ratings. It began to affect perceptions of other related political issues. One example is perceptions of the honesty of the election. Historically, both Muslims and Christians in Nigeria have held similar views on this issue, with trust in the electoral process consistently low. In 2011, for example, both groups showed nearly identical levels of trust, with 21% of Muslims and 21% of Christians believing in election honesty. This common trajectory even saw a simultaneous jump in 2011, with trust levels rising to 54% among Muslims and 51% among Christians.

After Buhari's election, however, this shared perspective began to diverge. By 2017, a significant gap had emerged: 69% of Muslims believed in the honesty of the elections, compared to just 29% of Christians. This shift underscores the profound impact that a polarizing figure or event can have on a range of interrelated issues in public discourse. Not only is the primary issue polarized, but the spillover effects can alter perceptions and beliefs on a host of related issues.

These temporal dynamics of the breakdown of consensus or neutrality, as observed in various global contexts, are vividly reflected in the digital realm of online social networks. A Japanese study of vaccination positions on Twitter illustrates the dissolution of a neutral state [104]. In the early stages of the discussion, the vast majority of users expressed neutral or ambivalent views about COVID-19 vaccination, suggesting that digital spaces do not necessarily start with clear partisan boundaries. Over time, however, this neutral landscape underwent a significant transformation. The dominant neutral voices gradually receded, overshadowed by a rising tide of pro-vaccine sentiment. This shift was influenced by key user accounts, information sources, and the interconnected web of responses within the Twitter community.

Such findings underscore the fluidity and responsiveness of online discussions. While real-world debates may take years or even decades to polarize, online spaces can reflect these shifts in a condensed timeframe, sometimes within weeks or months. This rapid evolution underscores the importance of online social networks as both a barometer for gauging public sentiment and a crucible in which that sentiment is shaped and reformed.

In the vast expanse of digital discourse, the flow of public sentiment can be both swift and profound. A study of multiple Twitter datasets across several countries, covering topics such as #MeToo, #Immigrants, #ClimateChange, and especially #Coronavirus, provides a compelling illustration of this dynamic [105]. For example, the initial discourse surrounding #Coronavirus on March 12, 2020 was characterized by a prevailing positive sentiment. But just one month later, on April 12, 2020, the landscape had changed dramatically. Hate speech began to eclipse positive discussions, marking a shift from a deliberative, consensual discourse to one characterized by division and hostility.

This shift in the digital conversation, from consensus to division, is emblematic of the inherent volatility of online spaces. Unlike traditional debates, which can evolve over long periods of time, the digital realm can encapsulate these shifts in sentiment in a remarkably condensed timeframe. Such rapid transitions are influenced by a confluence of factors: the influx of new information, the resonance of influential voices, and the intricate interplay of user interactions within the network.

But even in scenarios where rapid partisan delineation is evident, such as the COVID-19 debate in the United States, certain patterns emerge that challenge our understanding of online polarization. A study that examined Twitter discourse from January 2020 to April 2020 provides a compelling illustration [106]. The researchers used the ratio of intra- and inter-state communication as an indicator of partisanship, with higher values indicating a localized restriction of communication channels.

While the study observed a general trend of increasing compartmentalization over time, two key moments stood out. First, on January 22, as the debate was just beginning, a significant proportion of tweets originated from Washington DC, effectively breaking the news. This resulted in a significant cross-communication ratio. Later, on February 29, a similar phenomenon was observed when the first COVID-19 death was reported in the US. This critical event temporarily dissolved the previously established state-centric divide. In the aftermath, however, the divide not only re-emerged, but intensified.

Such observations underscore the potential impact of critical events on the trajectory of online debates. While these events may temporarily blur the lines of partisanship due to the rapid spread of information across boundaries, they can also shift communication patterns to resemble the unaligned scenario we propose. This suggests that even in deeply divided debates, specific events can temporarily obscure the overarching narrative.

This has two crucial implications: (1) The specific point in time from which a discussion is analyzed, essentially the initial state of a graph in our model, becomes paramount. This is because the nature of the discourse can be significantly influenced by external events that may temporarily alter the prevailing sentiment and communication patterns. (2) A model that aims to understand social polarization from social network graph data alone must be versatile enough to accommodate both unaligned and partisan situations. The former essentially represents the more challenging aspect of polarization, where the lines of division are not immediately apparent.

Appendix D. Extended synthetic evaluation

D.1. Scenario 1 — external connectivity for 2 communities

In the first additional scenario, we want to investigate the effects of the integration strength in the case of two communities. For this purpose, we define two communities $\mathcal{V}_c = \{c_1, c_2\}$ with $|\mathcal{V}_u[c_1]| = |\mathcal{V}_u[c_2]| = 2,500$ and 40 gatekeepers each. In addition to the human decision function and recommendation strategy, we take external connectivity with $C = \{0.3, 0.2, 0.1\}$ as the independent variable yielding a $\mathcal{L} \times \mathcal{A} \times C$ study setup. Compare Fig. F.5(a) to F.5(c) for a visualization of the initial graphs.

D.2. Scenario 2 — external connectivity for 3 communities

Under the second scenario, we examine the integration strength for three communities. We define three communities $\mathcal{V}_c = \{c_0, c_1, c_2\}$ where $\mathcal{L}_{c_0} = \mathcal{L}_1$ and $\mathcal{L}_{c_1} = \mathcal{L}_{c_2} = \mathcal{L}_2$. We set $|\mathcal{V}_u[c_0]| = 4,000$ (with 100 gatekeepers) and $|\mathcal{V}_u[c_1]| = |\mathcal{V}_u[c_2]| = 500$ (with 10 gatekeepers each). Again, this results in $\mathcal{L} \times \mathcal{A} \times C$ setup, where we set $C = \{0.4, 0.3, 0.2\}$ for the fringe communities.

D.3. Scenario 3 — size of 3 communities

In the third scenario, we want to look at how the size of the marginal communities affects social polarization. To do this, we vary their size with $S = \{1000, 1500, 2000\}$, while adjusting the center community's accordingly and keeping external connectivity constant at 0.9. Thus we have a $\mathcal{L} \times \mathcal{A} \times S$ setup.

D.4. Results

The results of our additional evaluations are essentially consistent with those from our main investigation in Section 4. For the first scenario, we continue to observe that the two conditions following \mathcal{L}_2 quickly reach high values for latent polarization, modularity and homophily for any degree of integration Fig. F.5. This is obvious in that the degree of integration in the scenarios shown here is set higher than before. Thus, the results underline that shared group identity robustly leads to fragmentation of the debate space in the presence of initial group affiliation. The condition $\mathcal{L}_1 \times Random$ also shows comparable behavior to before, as it quickly lapses into divided consensus, with the number of epochs required for this to happen becoming somewhat larger as initial separation increases. The results for condition $\mathcal{L}_1 \times Random$ are more difficult to interpret since the metrics for 0.1 are quite comparable to the \mathcal{L}_2 conditions while the latent polarization for the integration degrees 0.3 and 0.2 drops notably. This finding can be explained by the fact that the separation is already clearly pronounced at the beginning and consequently also transfers to the positioning in the latent information space. With regard to the graph-based question segmentation measures, however, there is a recognizable discrepancy. The decomposition in the latent representations only partially transfers to the connectivity of the graph. A progressive increase in polarization is thus still only achieved with decision function \mathcal{L}_2 .

With regard to Scenario 2, we consistently find the same results as in the main study. The degree of interconnectivity thus hardly influences how the individual conditions behave Fig. F.6. This is remarkable in that condition $\mathcal{L}_2 \times Personalized$ apparently still leads to polarization tendencies even with weaker integration. Consequently, personalization and social group behavior bridge the weakening of linkage and consistently ensure the formation of consistent group integrity. For condition $\mathcal{L}_1 \times Personalized$, we document moderate polarization values, which, however, only manifest themselves in the fact that the merging of communities drags on.

Finally, Scenario 3 shows that the size of the marginal communities is essential to whether a community retains its integrity under a broad recommendation spectrum Fig. F.7. Compared to the initial graph in Section 4, the two fringe communities proportionally occupy a larger space, resulting in condition $\mathcal{L}_2 \times Random$ also showing progressive moderate polarization tendencies. A crucial question for the significance of recommendation technology is therefore to what extent it contributes to the congruence of initially small communities whose ideas then may increasingly spread into the mainstream. This is especially important when a group of people specifically tries to popularize certain views. Our results suggest that this is difficult to achieve if, firstly, there is not a sufficient amount of potential amplifiers available or, secondly, algorithmic dissemination does not allow for it.

Appendix E. Latitude of acceptance

The main independent variables in each of the synthetic scenarios now to be presented are, first, the human decision function $\mathcal{L} \in \{\mathcal{L}_1, \mathcal{L}_2\}$ and, second, the choice of recommendation algorithm $\mathcal{A} \in \{\text{personalized}, \text{random}\}$. By random, we mean that message recommendations are not calculated with respect to some personalization logic, but are rather chosen arbitrarily from any graph user. Since the most important hyperparameter of our procedure is the latitude of acceptance, we define $\lambda \in \{0.1, 0.2, \dots, 0.9\}$ as an additional independent variable yielding a three-fold $\mathcal{L} \times \mathcal{A} \times \lambda$ study setup. Furthermore, we fix both the regular as well as the ideological sharpness parameter to $\mu = \kappa = 5$. Furthermore, we expect a regular member of both communities to have an average of $\psi_c^m = 5$ users connected to them. In case of gatekeepers, we assume $\psi_c^g = 20$ connections. For all users, we set $\psi = 5$. Finally, we set the number of message recommendations to 20 per epoch per user.

E.1. Scenario 1 — echo chamber

In our first scenario, we aim to approach the phenomenon of echo chambers which we define as a (small) subset of nodes that collectively separate themselves communicatively from the dominant standpoint by coupling closely together while, at the same time, disclosing cross-cutting communication channels to a large degree. We begin by defining $\mathcal{V}_c = \{c_h, c_e\}$ to consist of two communities, where c_h is the hegemonic community and c_e depicts the echo chamber respectively. Since we assume hegemonic dominance, i.e. $|\mathcal{V}_u[c_e]| \ll |\mathcal{V}_u[c_h]|$, we set $|\mathcal{V}_u[c_h]| = 1000$ (with 10 gatekeepers) and $|\mathcal{V}_u[c_e]| = 100$ (with a single gatekeeper). Finally, we define $\mathbf{p}_{c_h} = (0.95, 0.05)$ and $\mathbf{p}_{c_e} = (0.2, 0.8)$ to produce rather polarized but not completely separated communities. The resulting user graph consists of 7,394 edges. The corresponding communication graph contains an additional 4,224 message nodes and a total of 27,523 edges.

E.2. Scenario 2 — two-sided polarization

In our second setup, we aim to model the process of two larger communities competing against each other. We assume both communities to exhibit pronounced homophily while still being intertwined. Formally, we again define two communities $\mathcal{V}_c = \{c_1, c_2\}$ with $|\mathcal{V}_u[c_1]| = |\mathcal{V}_u[c_2]| = 500$ (10 gatekeepers each). Following the general rule of thumb of 30% of network communication being cross-cutting [8,14], we define $\mathbf{p}_{c_1} = (0.7, 0.3)$ and $\mathbf{p}_{c_2} = (0.3, 0.7)$ respectively. The user graph consists of 6,693 edges. For the communication graph, we append 3,876 message nodes and observe 25,151 edges.

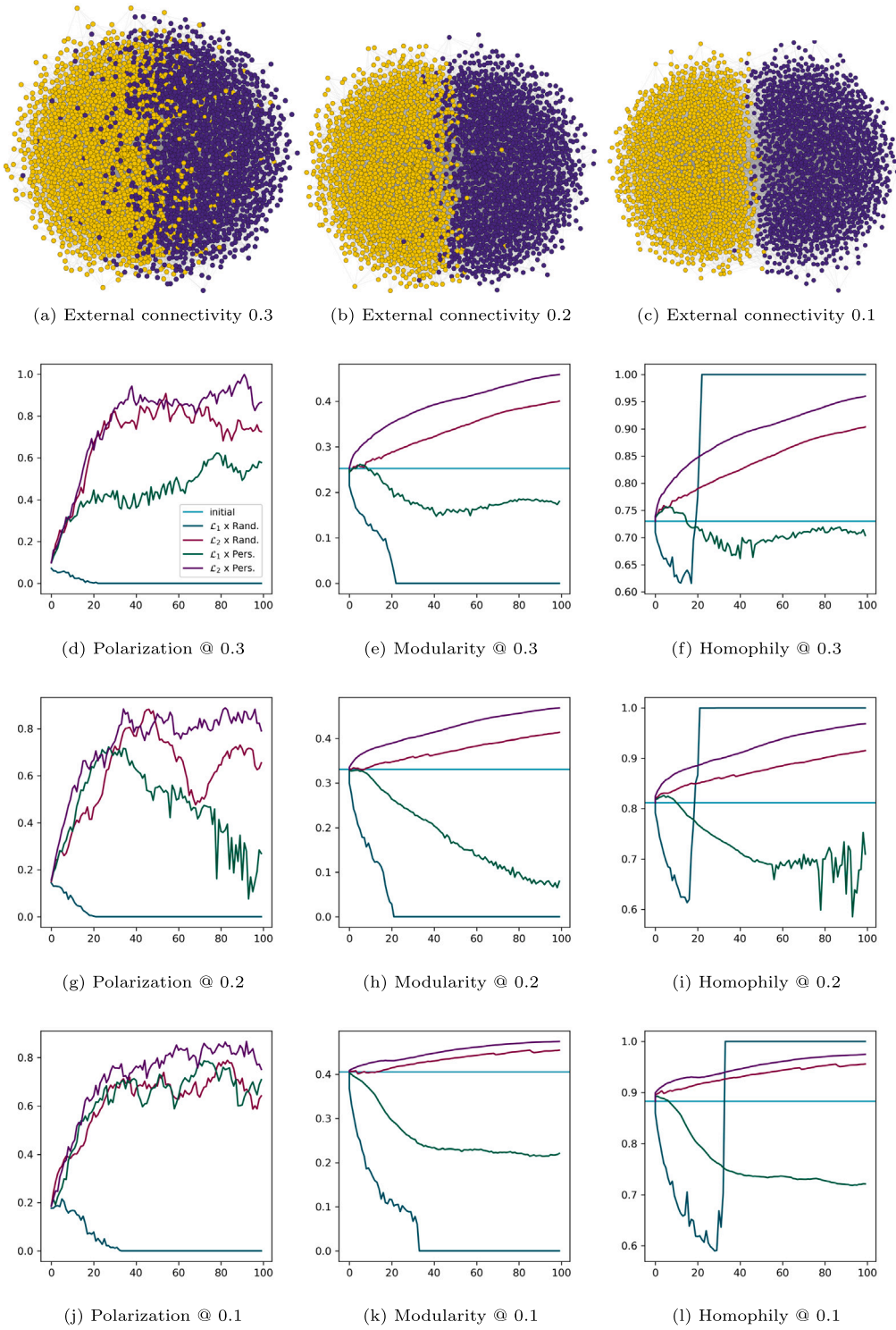


Fig. F.5. Initial graph topologies and metrics for the additional scenario of 2 communities. The graphs differ in terms of integration strength between the communities.

E.3. Scenario 3 — two-sided polarization (with an unaligned community)

In the third scenario, we want to inspect the case where a large proportion of the population is not yet leaning towards a specific standpoint. Oftentimes, especially in case a topic is highly ambiguous, many people are undecided with which side of a debate to solidarize while others quickly adopt a standpoint that aligns with the view of their ideological group. Hence, we propose the community of undecided

interlocutors to follow \mathcal{L}_1 until they have decided for a standpoint while members of the two polarized communities apply \mathcal{L}_2 .

Formally, we define three communities $\mathcal{V}_c = \{c_0, c_1, c_2\}$ where $\mathcal{L}_{c_0} = \mathcal{L}_1$ and $\mathcal{L}_{c_1} = \mathcal{L}_{c_2} = \mathcal{L}_2$. We set $|\mathcal{V}_u[c_0]| = 800$ (with 10 gatekeepers) and $|\mathcal{V}_u[c_1]| = |\mathcal{V}_u[c_2]| = 100$ (with a single gatekeeper each). Concerning inter-community connectivity, we let c_1 and c_2 be entirely separated from each other. Still, both produce some edges towards the undecided community c_0 and vice versa. Accordingly, $\mathbf{p}_{c_0} = (0.8, 0.1, 0.1)$ as well as $\mathbf{p}_{c_1} = (0.25, 0.75, 0.0)$ and $\mathbf{p}_{c_2} = (0.25, 0.0, 0.75)$. The user graph consists

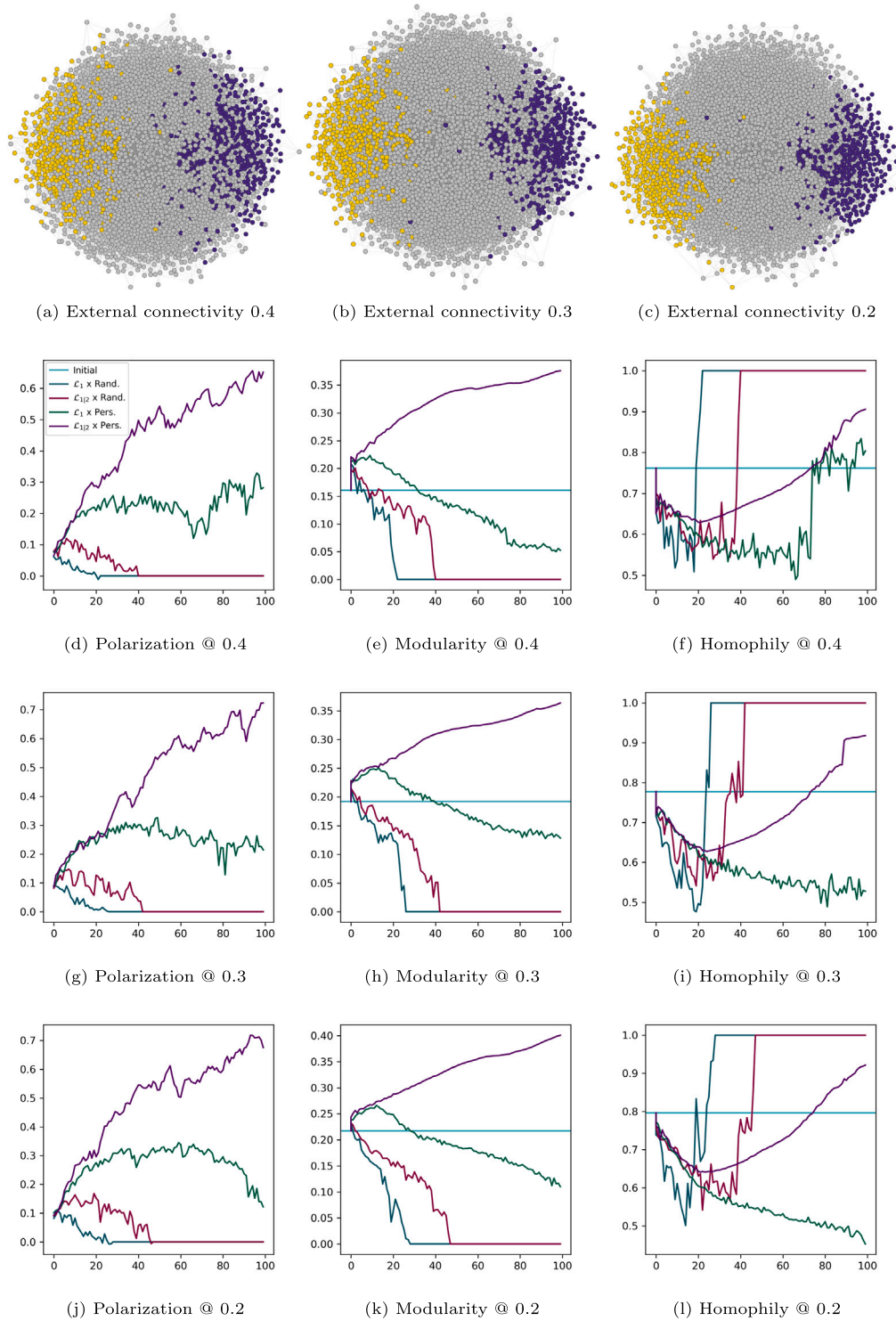


Fig. F.6. Initial graph topologies and metrics for the additional scenario of 3 communities. The graphs differ in terms of integration strength between the communities.

of 6,632 edges and the communication graph has 3,813 message nodes and a total of 24,737 edges.

E.4. Results

The results are depicted in Figs. F.8–F.10. First, we note that, as expected, increasing the latitude of acceptance leads to increased network communication, which is primarily manifested by a larger number of simulatively formed edges — this is generally true for all scenarios and combinations. Specifically, it can be observed that both

conditions based on decision rule \mathcal{L}_1 form significantly more edges than those following \mathcal{L}_2 across all scenarios.

Furthermore, the choice of recommendation algorithm apparently impacts edge formation as well. For example, even for a low latitude of acceptance, $\mathcal{L}_1 \times Random$ tends towards a common consensus, which manifests itself in the previously defined communities quickly coinciding. In contrast, personalized message recommendations preserve preferential structures shared between community members and may even reinforce them such that communities remain largely stable in both the first and second scenarios. The results are thus consistent with

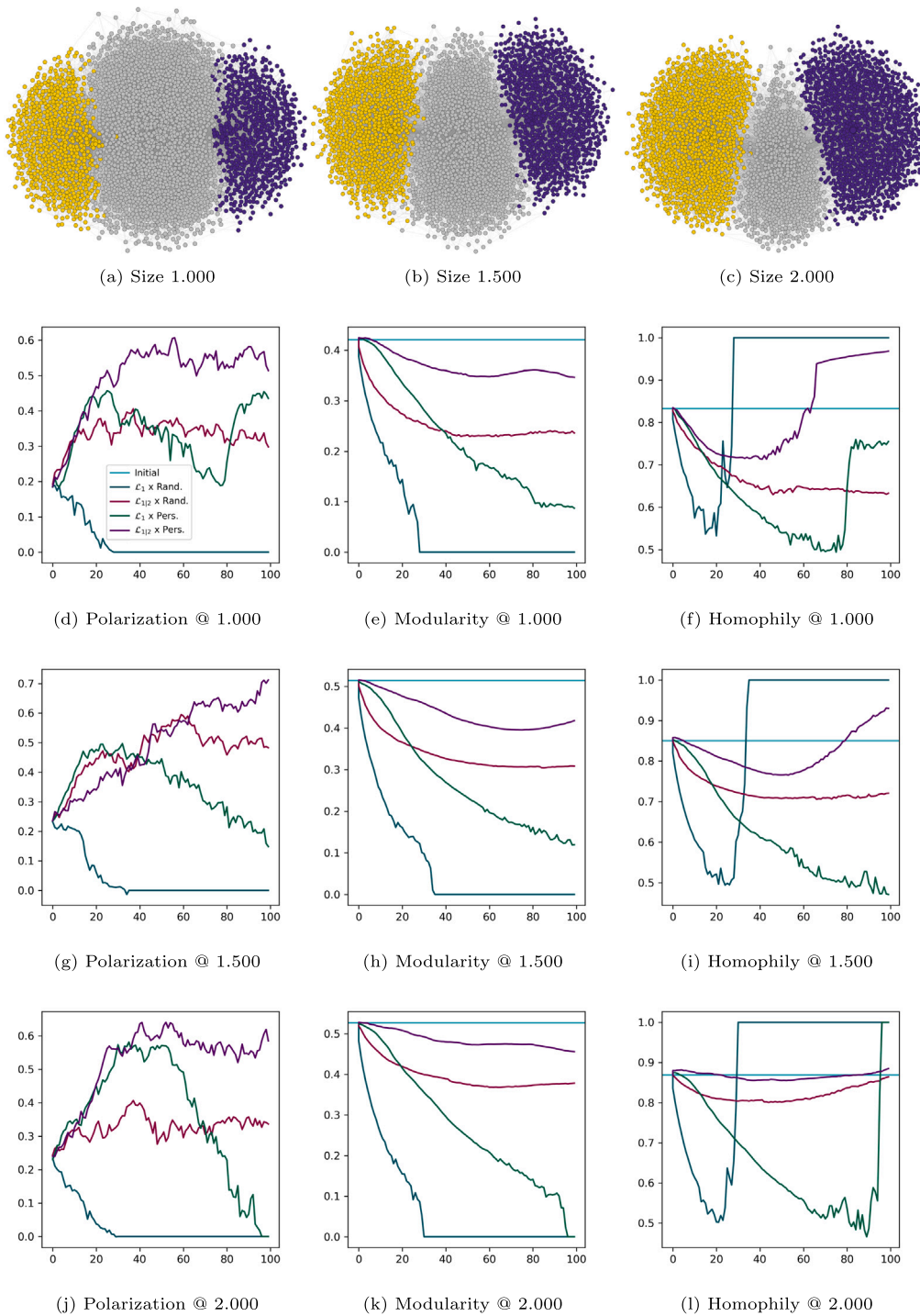


Fig. F.7. Initial graph topologies and metrics for the additional scenario of 3 communities. The graphs differ in terms of community sizes.

the intuition of filter bubbles [6], according to which personalized content selection leads to condensed communication channels preventing a shared consensus from materializing. As the first scenario shows, these information horizons can be so limited that even a small community that deviates from the hegemonic viewpoint is able to stabilize itself.

However, social polarization not only implies the stabilization of communities, but also mutual radicalization and, thus, a progressive distancing from consensus. In the case of condition $\mathcal{L}_1 \times Personalized$, however, this centrifugal movement is actually not observed. Increased

polarization potentials according to homophily and modularity can be identified exclusively for conditions that follow \mathcal{L}_2 . Especially in the case of two-sided polarization (scenario 2), we can see that the respective conditions achieve a significantly higher differentiation potential between the two communities, which is manifested in particular by the fact that the homophily and modularity values of the initial network are even increased.

The progressive delineation of two viewpoints by the application of \mathcal{L}_2 , whether in an equally weighted bipartite scenario or with respect

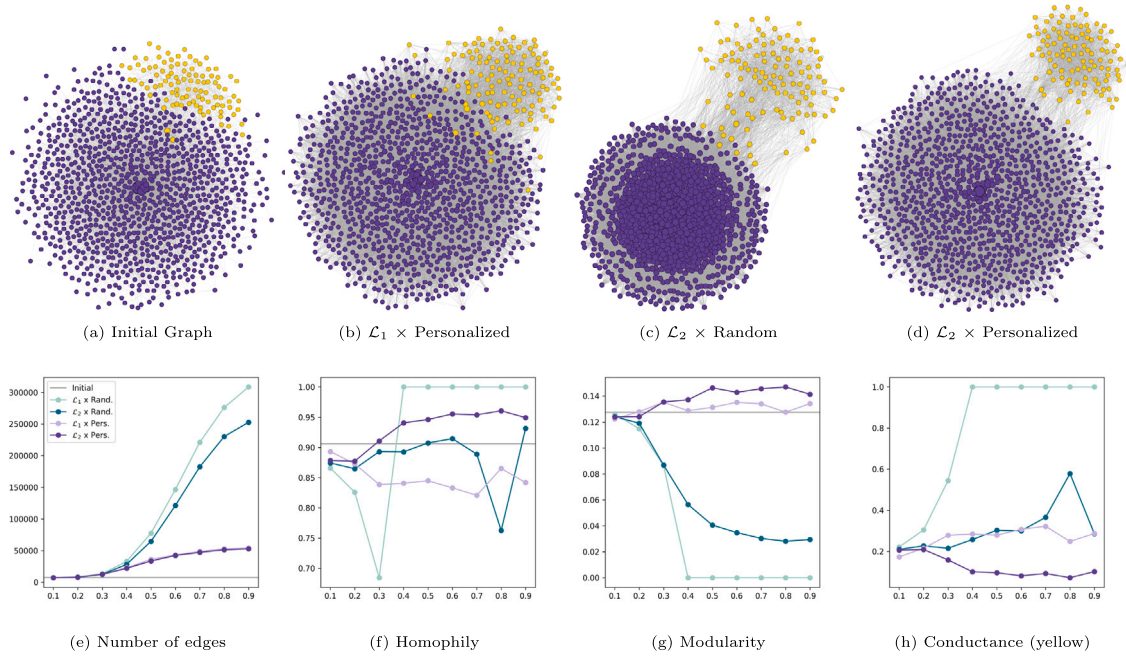


Fig. F.8. Scenario 1 — echo chamber: graph topology & metrics.

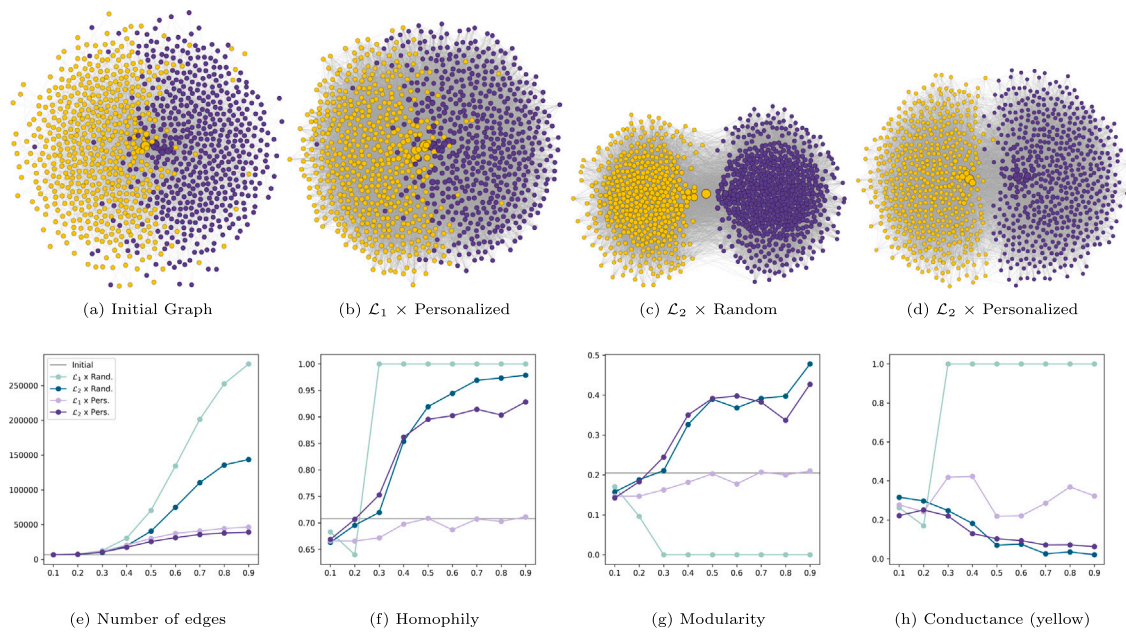


Fig. F.9. Scenario 2 — two-sided polarization: graph topology & metrics.

to a small echo chamber, can also be verified by inspecting the graph topology. For example, the echo chamber shown in yellow is moving away from consensus, allowing for increasingly clear delineation. It should be noted that this movement occurs regardless of the recommendation algorithm used. External sources are accordingly excluded to a reasonable extent to ensure the dynamics of one’s group.

What stands out when inspecting the graph topology with respect to the recommendation algorithm is the fact that random recommendations promote egalitarian exchange. This shows primarily in that existing opinion leaders in the original graph, while not eliminated entirely, are at least hemmed into a more balanced exchange. This is due to the fact that, overall, there is significantly stronger networking within the communities (especially in the dominant one), which also explains the possibly initially surprising reduction in modularity for

$\mathcal{L}_2 \times Random$ in scenario 1. Opposed to this, the use of personalization seemingly comes closer to reality because for $\mathcal{L}_2 \times Personalized$ it is quite clear that opinion leaders can maintain their influence over time.

Finally, we assume that scenario 3 most closely reflects the evolution of a debate on real social networks, since many participants often join a position only in the course of the exchange. Accordingly, we consider the community-dependent assignment of decision rules \mathcal{L}_1 and \mathcal{L}_2 to be reasonable — at least if the latitude of acceptance is chosen adequately. The most stable results in this regard are observed with values between 0.4 and 0.7. With respect to the resulting graphs and community affiliation, this scenario differs from the second one quite clearly in that $\mathcal{L}_1 \times Personalized$ cannot produce a bipartite polarization, whereas in the second scenario this was designed in advance. The ideological movement of people, i.e., their change of opinion or

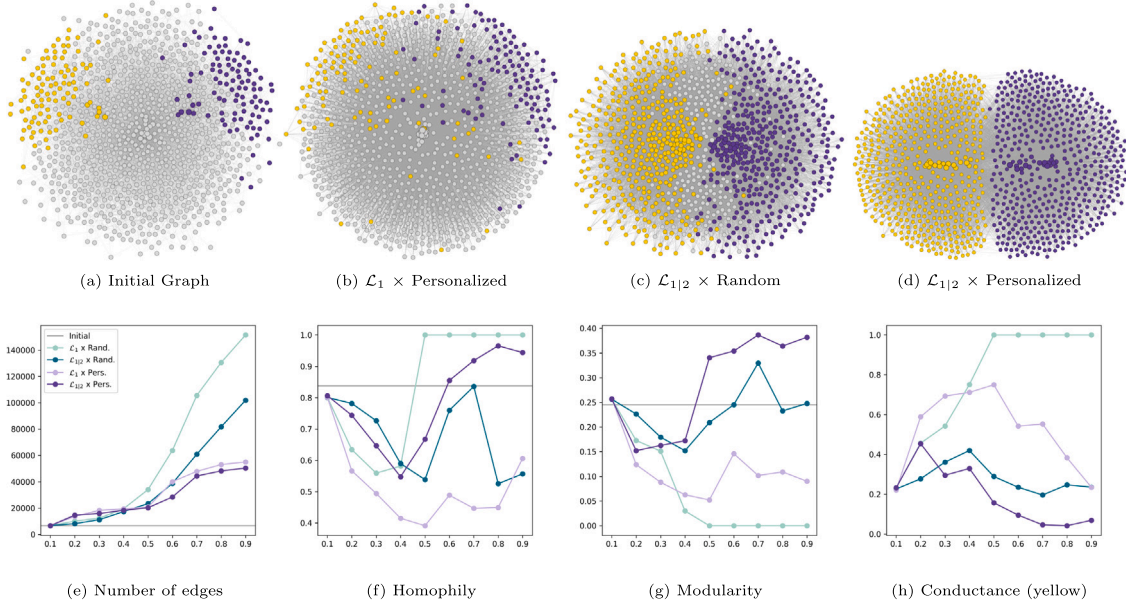


Fig. F.10. Scenario 3 — two-sided polarization (with an unaligned community): graph topology & metrics.

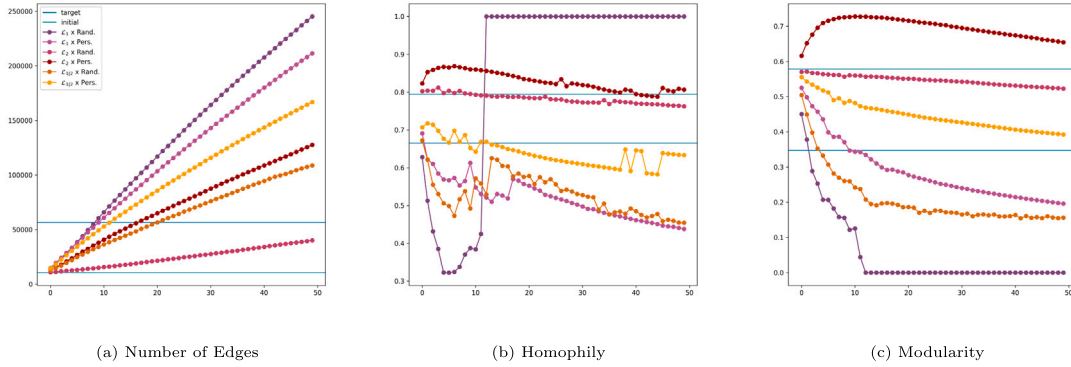


Fig. F.11. Real-world dataset: metrics over 50 epochs.

commitment to a point of view, is thus not simulated adequately at all in the second scenario. In this sense, scenario 3 particularly illustrates the consideration of psychological group effects. It is only through intra-group cohesion that it is possible to spread one’s own point of view.

Moreover, the results indicate that the use of personalization can further accelerate such polarization tendencies. In this sense, our method realizes the reciprocal nature of the relationship between group dynamics and recommendation personalization that we have hypothesized.

Appendix F. Real-world graph evaluation

In this section, we want to demonstrate that our model is capable of handling real-world data as well. We have gathered a communication graph via the Twitter API concerning the topic *Ukraine* from 01. November 2021 to 23. February 2022. Since our method is sensitive to data sparsity, we have densified the collected graph to meet the requirements. Concretely, we require a minimum of 20 Retweets per user, 20 Retweets per Tweet as well as 5 Tweets per author user (that is, the number of messages sent is either 0 or greater or equal to 5). We define the resulting communication graph consisting of 4,946 users, 2,225 messages and 75,970 edges as our target graph \mathcal{G}_c^t (see Fig. 4). The objective in this scenario is to approximate the community

structure of the target user graph \mathcal{G}_u^t as closely as possible given an initial subgraph \mathcal{G}_c^0 which we define to consist of the 20% earliest edges found in \mathcal{G}_c^t (while ensuring a connected component).

Concerning independent variables, we again choose human decision function and recommendation algorithm. For the human decision function, we select from $\mathcal{L} \in \{\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_{1|2}\}$, where $\mathcal{L}_{1|2}$ means that we heuristically choose the $k = 5$ most homophilic communities to exhibit group-referential behavior while the remaining do not. Please note that the selection of k is a fundamental parameter to the performance of condition $\mathcal{L}_{1|2}$ and is not trivial to assign. Depending on the given graph structure, selecting k might involve human intervention, for instance via manual inspection of the data.

Weak-tie recommendations. We choose the recommendation algorithm again as $\mathcal{A} \in \{\text{personalized}, \text{random}\}$. In this scenario, however, we extend the recommendation pipeline by considering *weak-tie* recommendations. Weak-ties play an essential role in how network communication is structured on modern social media platforms as they depict a means of how previously disconnected users may get in touch [107,108]. A prominent example of weak-tie recommendations is Twitter’s retweet functionality, as retweets by close-tie users are often displayed in a user’s timeline annotated as “Close-tie user u has retweeted message m ” giving the target user the possibility to categorize a recommendation accordingly. Formally, we extend the set of message candidates C_m to not only include messages sent by candidate user U ,

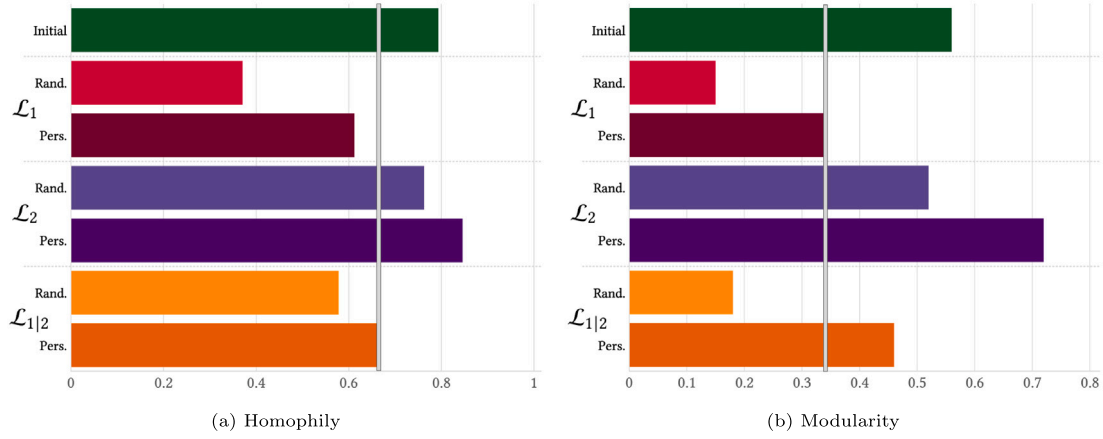


Fig. F.12. Real-world Dataset: Homophily & Modularity at epoch where the difference of the number of edges to the target graph is smallest. The vertical line depicts the target graph's value.

but also messages they themselves have reacted to in the past. Hence, we redefine C_m as:

$$C_m = \{m | \exists U : \tau_s(m, U) = 1 \vee \tau_r(U, m) = 1 \wedge U \in C_u\} \quad (\text{F.1})$$

In terms of logical queries in latent embedding space, searching for weak-tie message recommendations corresponds to the following query:

$$M_\gamma, \exists M, U : e_1 \wedge e_2 \wedge e_3, \quad (\text{F.2})$$

$$\text{where } e_1 = \tau_r(v, M) \wedge e_2 = \tau_s(M, U) \wedge e_3 = \tau_3(U, M_\gamma)$$

$$\wedge v, U \in \mathcal{V}_u \wedge M_\gamma, M \in \mathcal{V}_m \wedge \tau_r, \tau_s \in \mathcal{R}$$

Results. After verifying that our proposed method is in principle capable of simulating the progressive delineation of (ideologically shaped) groups on synthetic data, we now want to address the question to which extent this capability can also be transferred to real-world datasets. Comparing the empirical graph with the synthetic ones, we first note that the prior obviously contains much more pronounced subnetwork structures. The community detection algorithm [71] thus identifies 21 sub-structures for the initial graph. Although there are obvious tendencies towards the consolidation of information channels, the topology at the beginning of the debate appears broad and highly differentiated. In comparison, the final target graph shows more clearly delineated communities, two of which in particular (shown in orange and yellow) stand out distinctly from the central discourse. Overall, the number of communities has decreased significantly to 5.

The quality of a simulation condition is measured with respect to how well this network structure is approximated with respect to our already introduced metrics, i.e. in this case homophily and modularity. In general, we find that methods following decision rule \mathcal{L}_1 significantly underestimate the fragmentation of the target graph, while the reverse is true in the case of \mathcal{L}_2 (compare Fig. F.11). Specifically, we find that both personalization of recommendations and group-referential decision behavior favor the preservation as well as the generation of homophilic structures. Whereas $\mathcal{L}_1 \times \text{Random}$ tends to lapse into consensus very quickly, $\mathcal{L}_1 \times \text{Personalized}$ is able to maintain some structural stability, at least for a while. In contrast, the condition $\mathcal{L}_2 \times \text{Personalized}$ ensures that fragmentation tendencies are very clearly overemphasized, which even leads to an increase in modularity compared to the initial network. Coupling \mathcal{L}_2 with random recommendations at least reverses the increase in fragmentation, but not to an extent to speak of a reasonable approximation to the target network. Hence, it is obviously counterproductive to assign decision rule \mathcal{L}_2 to each identified community, since in this case the degree of fragmentation is significantly overestimated. Rather, over time, different sub-communities appear to coalesce into larger ones, which is why a combination of both human

decision rules ($\mathcal{L}_{1|2} \times \text{Personalized}$) produces the most stable approximation. Overall, we hold that we cannot achieve good prediction of network communication solely through \mathcal{L}_2 , nor through personalized content selection alone. Rather, it seems that both must be present to reciprocally influence each other.

In order to verify our findings in more detail, we next look at the concrete network state at the point in time when the simulation procedure of a condition shows the smallest difference with respect to the number of formed edges to the target network (compare Figs. 4 and F.12). With respect to homophily and modularity, it is still evident here that \mathcal{L}_1 tends to underestimate fragmentation, while \mathcal{L}_2 overemphasizes it. Looking at the network topology of different conditions provides especially interesting insights: $\mathcal{L}_1 \times \text{Personalized}$ is in principle able to identify fragmentation tendencies, but overall the extent seems to be too low. It should be noted, however, that the two more clearly delineated communities (orange and yellow) were identified in this case. With respect to the general structure of the network, the $\mathcal{L}_2 \times \text{Random}$ condition interestingly achieves a good approximation. While the overall degree of fragmentation is clearly too high, finer details of the target network, however, are positively captured. Finally, $\mathcal{L}_{1|2} \times \text{Personalized}$ provides a reasonable level of fragmentation. Both partially-separated communities are clearly identifiable. However, some structural details are lost that are still preserved in the case of \mathcal{L}_2 . We take this as motivation to apply hierarchical community structure in the future to represent sub-community structures as well.

References

- [1] M. Prior, Media and political polarization, *Annu. Rev. Political Sci.* 16 (1) (2013) 101–127, <https://doi.org/10.1146/annurev-polisci-100711-135242>, Publisher: Annual Reviews.
- [2] Y. Lelkes, G. Sood, S. Iyengar, The hostile audience: the effect of access to broadband internet on partisan affect, *Am. J. Political Sci.* 61 (1) (2017) 5–20, <https://doi.org/10.1111/ajps.12237>, Publisher: Wiley Online Library.
- [3] P. Törnberg, C. Andersson, K. Lindgren, S. Banisch, Modeling the emergence of affective polarization in the social media society, *PLoS One* 16 (10) (2021) e0258259, <https://doi.org/10.1371/journal.pone.0258259>, Publisher: Public Library of Science San Francisco, CA USA.
- [4] D.C. Mutz, *Hearing the Other Side: Deliberative Versus Participatory Democracy*, Cambridge University Press, 2006, <https://doi.org/10.1017/CBO9780511617201>.
- [5] H. Lee, N. Kwak, S.W. Campbell, Hearing the other side revisited: the joint workings of cross-cutting discussion and strong tie homogeneity in facilitating deliberative and participatory democracy, *Commun. Res.* 42 (4) (2015) 569–596, <https://doi.org/10.1177/009365021348382>, Publisher: SAGE Publications Sage CA: Los Angeles, CA.
- [6] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*, Penguin Group, The, ISBN: 1-59420-300-8, 2011.

- [106] J. Jiang, E. Chen, S. Yan, K. Lerman, E. Ferrara, Political polarization drives online conversations about COVID-19 in the United States, *Hum. Behav. Emerg. Technol.* 2 (3) (2020) 200–211, <https://doi.org/10.1002/hbe2.202>, Publisher: Wiley Online Library.
- [107] L. Weng, M. Karsai, N. Perra, F. Menczer, A. Flammini, Attention on weak ties in social and communication networks, in: *Complex Spreading Phenomena in Social Systems*, Springer, 2018, pp. 213–228, https://doi.org/10.1007/978-3-319-77332-2_12.
- [108] J. Zhao, J. Wu, K. Xu, Weak ties: Subtle role of information diffusion in online social networks, *Phys. Rev. E* 82 (1) (2010) 016105, <https://doi.org/10.1103/PhysRevE.82.016105>, Publisher: APS.

DuEPublico

Duisburg-Essen Publications online

UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

ub | universitäts
bibliothek

Diese Dissertation wird via DuEPublico, dem Dokumenten- und Publikationsserver der Universität Duisburg-Essen, zur Verfügung gestellt und liegt auch als Print-Version vor.

DOI: 10.17185/duepublico/81491

URN: urn:nbn:de:hbz:465-20240207-124410-5

Alle Rechte vorbehalten.