Investigating Political and Informational Homogeneity in Social Media Using Computational Methods

—

Untersuchung der politischen und informationellen Homogenität in sozialen Medien mit Hilfe von computergestützten Methoden

Von der Fakultät für Ingenieurwissenschaften,
Abteilung Informatik und Angewandte Kognitionswissenschaft
der Universität Duisburg-Essen

zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte kumulative Dissertation

von
Daniel Röchert
aus
Moers

Gutachter: Prof. Dr. Stefan Stieglitz
Gutachter: Jun.-Prof. Dr. German Neubaum

Tag der mündlichen Prüfung: 20.06.2022

“I did not direct my life. I didn't design it. I never made decisions.
Things always came up and made them for me. That's what life is.”

— B. F. Skinner

Acknowledgements

This dissertation was written during my work as a research associate in the junior research group Digital Citizenship in Network Technologies (DICINT) at the University of Duisburg-Essen.

First of all, I would like to express my special thanks to Jun.-Prof. Dr. German Neubaum, in particular for supervising my PhD project and his versatile and helpful support, without which it would not have been possible to complete this work. The interdisciplinary work between social psychology and computer science in the junior research group not only broadened my horizons and gave these individual sciences greater insight into the work of the other, but also led to exciting discussions on a variety of subjects.

I would also like to thank my dissertation advisor Prof. Dr. Stieglitz for the numerous discussions and the professional debate, which helped me greatly in completing the dissertation.

Furthermore, I thank my companion and friend Manuel Cargnino, with whom I started the PhD journey together in 2018. During this time, we have not only supported each other, but have also shared a variety of experiences that have bonded us together.

I would also like to thank the fellow members of the Digital Communication and Transformation (digicat) department, who integrated me into their team and thus made an important contribution.

Finally, my special thanks go to my family Miriam, Kerstin, and Andreas Röchert, in addition to many friends, who made this path possible for me and supported and encouraged me during my PhD project, as well as in the up and down phase

Research papers included in the cumulus

Paper 1: Röchert, D., Neubaum, G., & Stieglitz, S. (2020). Identifying Political Sentiments on YouTube: A Systematic Comparison Regarding the Accuracy of Recurrent Neural Network and Machine Learning Models. In *Multidisciplinary International Symposium on Disinformation in Open Online Media* (pp. 107-121). https://doi.org/10.1007/978-3-030-61841-4_8

Paper 2: Röchert, D., Neubaum, G., Ross, B., Brachten, F., & Stieglitz, S. (2020). Opinion-based homogeneity on YouTube: Combining Sentiment and Social Network Analysis. *Computational Communication Research*, *2*(1), 81-108. https://doi.org/10.5117/CCR2020.1.004.ROCH

Paper 3: Röchert, D., Neubaum, G., Ross, B., & Stieglitz, S. (2022). Caught in a networked collusion? Homogeneity in conspiracy-related discussion networks on YouTube. *Information Systems*, *103*, 101866. https://doi.org/10.1016/j.is.2021.101866

Paper 4: Röchert, D., Shahi, G. K., Neubaum, G., Ross, B., & Stieglitz, S. (2021). The Networked Context of COVID-19 Misinformation: Informational Homogeneity on YouTube at the Beginning of the Pandemic. *Online Social Networks and Media*, *26*, 100164. https://doi.org/10.1016/j.osnem.2021.100164

Paper 5: Röchert, D., Weitzel, M., & Ross, B. (2020). The homogeneity of right-wing populist and radical content in YouTube recommendations. In *International Conference on Social Media and Society* (pp. 245-254). https://doi.org/10.1145/3400806.3400835

Paper 6: Röchert, D., Cargnino, M., & Neubaum, G. (2022). Two sides of the same leader: an agent-based model to analyze the effect of ambivalent opinion leaders in social networks. *Journal of computational social science*, 1-47. https://doi.org/10.1007/s42001-022-00161-z

Paper 7: Cabrera, B., Ross, B., Röchert, D., Brünker, F., & Stieglitz, S. (2021). The influence of community structure on opinion expression: an agent-based model. *Journal of Business Economics*, *91*(9), 1331-1355. https://doi.org/10.1007/s11573-021-01064-7

Abstract

Social platforms such as YouTube, Facebook, and Twitter have become indispensable in today's world, as they provide a tool for communication and allow the exchange and dissemination of political opinions and information. In this context, the buzzword "echo chamber" is commonly used, as there is a concern that online social networks promote the confrontation of users with information and opinions that are in line with their own stance and beliefs. Accordingly, homogeneity in networks can also emerge among minorities communicating in a manner detached from majority societies and thus not having access to general discussion. In addition to the average users who use social media to discuss the news on a weekly basis, it is unknown at this point how homogeneous the communication networks are of marginalized groups that may be spreading extreme political views, misinformation, or conspiracy-theory content.

To investigate the complexity of opinions and information in terms of their homogeneity and to gain a deeper understanding of the underlying problems, the interdisciplinary research area of Computational Social Science (CSS) provides innovative computational methods to address social, political, and economic issues. The combination of social science theories with computer science methods makes it possible to develop new solutions to existing problems in order to explain the behavior of humans and their environment. The dissertation pursues the goal of applying a variety of methods of CSS integrated with Big Data datasets to identify patterns and relationships of human online communication and to provide an explanation for the emergence of political homogeneity in networks.

Accordingly, two fundamental aspects are explored in this dissertation. The first aspect relates to the prevalence of homogeneity of opinions and information on social platforms. Here, the concept of "opinion-based homogeneity" and "informational homogeneity" is presented, whereby a computational method combining natural language processing (NLP), machine learning (ML), and social network analysis (SNA) was developed to determine homogeneity in networks. For this purpose, four studies were designed that addressed a wide range of different topics (politically controversial as well as right-wing populist topics, conspiracy theories, and misinformation about COVID-19). The results indicate that political homogeneity is not widespread for the average user and that YouTube users engage in more heterogeneous communication behavior when discussing politically controversial topics or current political

situations (in the case of COVID-19). However, the results also highlight that marginalized groups of society, such as people advocating conspiracy theories, show a moderate level of homogeneity and are more likely to engage with like-minded people. The second aspect deals with the study of influencing factors that could be responsible for the emergence and change of homogeneity. To this end, two studies were designed using agent-based modeling to investigate, first, the influence of opinion leaders in networks and, second, the influence of online community structures on public opinion formation according to the assumptions of the spiral of silence theory. The results of both studies show that opinion leaders, as well as the structure of communities in networks, can be characterized as influencing factors for the change and emergence of political homogeneity, as they have an impact on temporal opinion formation.

In addition to theoretical implications, practical implications can also be derived from this dissertation. First, the dissertation offers a variety of methods that provide a blueprint for future research to determine homogeneity in networks. Second, the dissertation presents implications on how the insights gained can be used in the field of political marketing or political education. In summary, the dissertation contributes to the field of the emergence and change of political homogeneity in social networks and can contribute to a deeper understanding of this by exploring the prevalence of homogeneity and identifying influencing factors.

Zusammenfassung

Soziale Plattformen wie YouTube, Facebook und Twitter sind aus der heutigen Welt nicht mehr wegzudenken, da sie ein Kommunikationsmittel darstellen, welches für den Austausch und die Verbreitung von politischen Meinungen und Informationen dient. In diesem Zusammenhang wird häufig das Schlagwort "Echokammer" verwendet, da die Befürchtung besteht, dass soziale Online-Netzwerke homogene virtuelle Räume schaffen, in denen Nutzer nur mit Informationen und Meinungen konfrontiert werden, die mit ihren eigenen Ansichten und Überzeugungen übereinstimmen. Dementsprechend kann die Homogenität in Netzwerken auch dazu führen, dass Minderheiten losgelöst von der Mehrheitsgesellschaft kommunizieren und somit keinen Zugang zur allgemeinen Diskussion haben. Neben den Durchschnittsnutzern, welche soziale Medien wöchentlich zur Diskussion der Nachrichten nutzen, ist es zum derzeitigen Zeitpunkt unbekannt, wie homogen die Kommunikationsnetzwerke von Randgruppen sind, welche möglicherweise extreme politische Ansichten, Falschinformationen oder verschwörungstheoretische Inhalte verbreiten.

Um die Komplexität von Meinungen und Informationen auf deren Homogenität zu untersuchen und deren Problematik besser zu verstehen, kann der interdisziplinäre Forschungsbereich der Computational Social Science (CSS) behilflich sein, um mit innovativen computerbasierten methodischen Ansätzen, neue Lösungswege im sozialen, politischen und wirtschaftlichen Bereich zu finden. Die Kombination aus sozialwissenschaftlichen Theorien mit informatischen Methoden ermöglicht es, neue Lösungsansätze zu bisherigen Problematiken herauszuarbeiten, um das Verhalten von Menschen und ihrer Umwelt zu erklären. Dabei verfolgt die Dissertation das Ziel, eine Vielzahl von Verfahren der CSS in mit Big-Data-Datensätzen zu nutzen, um Muster und Zusammenhänge menschlicher Online-Kommunikation zu identifizieren und eine Erklärung für die Entstehung der politischen Homogenität in Netzwerken zu liefern.

In der vorliegenden Dissertation werden demnach zwei grundlegende Aspekte untersucht. Der erste Aspekt bezieht sich auf die Prävalenz der Homogenität von Meinungen und Informationen auf sozialen Plattformen. Hierbei wird das Konzept der „meinungsbasierten Homogenität“ und der „informationellen Homogenität“ vorgestellt, bei dem eine computergestützte Methodenkombination aus NLP, Machine-Learning und Netzwerkanalyse entwickelt wurde, um die Homogenität in Netzwerken festzustellen. Hierfür wurden vier Studien konzipiert, welche sich mit einem breiten Spektrum unterschiedlichster Themen (politische kontroverse, als auch rechtspopulistische Themen, Verschwörungstheorien und Fehlinformation zu COVID-

19) auseinandergesetzt haben. Die Ergebnisse zeigen, dass die politische Homogenität für den Durchschnittskonsumenten nicht weit verbreitet ist und das YouTube Nutzer ein heterogeneres Kommunikationsverhalten entwickeln, wenn sie sich über politisch kontroverse Themen oder aktuelle Informationen zur aktuellen politischen Lage austauschen (im Fall von COVID-19). Allerdings zeigen die Ergebnisse auch, dass Randgruppen der Gesellschaft wie beispielsweise, Menschen die Verschwörungstheorien befürworten, ein moderates Level an Homogenität aufweisen und eher mit Gleichgesinnten in Kontakt treten. Der zweite Aspekt befasst sich mit der Untersuchung von einflussreichen Faktoren, die für die Entstehung und für die Veränderung der Homogenität verantwortlich sein könnten. Hierzu wurden zwei Studien mittels agentenbasierter Modellierung konzipiert, welche zum einen den Einfluss von Meinungsführern in Netzwerken untersucht hat und zum anderen, welchen Einfluss die Strukturen von Online-Gemeinschaften auf die öffentliche Meinungsbildung gemäß den Annahmen der Theorie der Schweigespirale haben. Die Ergebnisse beider Studien zeigen, dass sowohl Meinungsführer als auch die Struktur von Gemeinschaften in Netzwerken, als Einflussfaktoren zur Veränderung und Entstehung der politischen Homogenität charakterisiert werden können, da die einen Einfluss auf das zeitliche Meinungsbild haben.

Neben den theoretischen Implikationen lassen sich ebenso praktische Implikationen in der vorliegenden Dissertation ableiten. Zum einen bietet die Arbeit eine Vielfalt von Methoden an, welche eine Blaupause für zukünftige Forschung bereitstellt, um die Homogenität in Netzwerken zu ermitteln. Zum anderen zeigt die Arbeit Möglichkeiten auf, wie die Erkenntnisse im Bereich des politischen Marketings oder der politischen Bildung genutzt werden können. Zusammenfassend leistet die Dissertation einen Beitrag im Bereich der Entstehung und Veränderung der politischen Homogenität in sozialen Netzwerken und kann durch die Erforschung der Prävalenz der Homogenität und der Identifizierung von Einflussfaktoren zu einem besseren Verständnis beitragen.

Table of Contents

Figures

Tables

Abbreviations

ABM	Agent-based modeling
BERT	Bidirectional encoder representations from transformers
CSS	Computational social science
DL	Deep learning
LR	Logistic Regression
ML	Machine learning
NLP	Natural language processing
SNS	Social networking sites
SNA	Social network analysis
SVM	Support vector machine
RNN	Recurrent neural network
TF-IDF	Term frequency–inverse document frequency
UGC	User-generated content

1. Introduction

1.1. Context and motivation

Social networking sites (SNS) are accused in the public domain of creating virtual spaces (so-called *echo chambers*) where people join homogeneous communities in which they are surrounded by like-minded people who reinforce their own stance in the form of ideology or opinions and might foster processes of polarization (Boutyline & Willer, 2017; Colleoni et al., 2014; Sunstein, 2017). In the field of social network analytics (SNA), this process is also referred to as homophily and describes "*the principle that a contact between similar people occurs at a higher rate than among dissimilar people*" (McPherson et al., 2001, p. 416). The literature on echo chambers and their existence in online social networks indicates controversy. Some researchers demonstrate the existence of potential echo chambers in online social networks (Cinelli et al., 2021; Colleoni et al., 2014; van Eck et al., 2021), while other researchers concede that there is no empirical evidence for the existence of homogeneous cocoons for ordinary people (Bruns, 2019a; Dubois & Blank, 2018). Not only the communication between liked-minded users, but also the use of SNS have raised concern that the consumption of personalized content, which is proposed on the basis of "*filter bubbles*" (Pariser, 2012) created by algorithms, can cause a distorted perception of public opinion (Neubaum & Krämer, 2017) and the formation of ideological homogeneous groups (Shah et al., 2017). The differing conclusions about the existence of filter bubbles and echo chambers might be caused by the lack of a clear scientific definition of the terms, leading to a problematic situation in which a wide range of interpretations is possible due to the fact that they represent a dichotomy (Bruns, 2019a). Therefore, this dissertation refers to the term of political homogeneity, which is a concept that refers to homogeneous thematic connections between political opinions and information in networks. Here, the question arises as to the dimensions in which this political homogeneity takes place, since communication on SNS can proceed in very different ways. On the one hand, political homogeneity can take place based on communication between individuals who exchange homogenous opinions and attitudes on political topics (opinion-based homogeneity); on the other hand, communication can also take place on the basis of information, whereby users can be influenced by homogenous and recurring homogeneous content and information from algorithms or misinformation (informational homogeneity).

The YouTube platform is not only currently the most widely used SNS (Auxier & Anderson, 2021; Newman et al., 2021), but also lends itself to the study of homogeneous spaces since it "*can engage users from opposite ideological camps through the distribution of controversial socio-political video content*" (Bliuc et al., 2020, p. 831) and foregrounds the political and social community aspect for communication in social networking platforms (Burgess & Green, 2018). YouTube also seems to be an ideal platform for investigating marginalized groups and their communication structures since, according to studies, YouTube's recommendation system seems to be susceptible to misinformation (Kaiser et al., 2021) and extremely politically radical content (Stöcker & Preuss, 2020). Moreover, the platform provides users with an opportunity to obtain information about alternative news (Newman et al., 2021). In general, there is the fundamental problem that the homogeneity of opinions and information might well lead to minorities communicating in a manner detached from a majority society, and thus not participate in the general discussion. This can lead to (a) an undermining of the diversity of opinion (Graham, 2015), (b) a narrowing of the (political) world view (Scheufele & Nisbet, 2013), and (c) polarization of opinions, which can lead to social fragmentation in which extreme opinions are represented (Bright, 2018; Prior, 2007; Sunstein, 2007). The high technological standard and ubiquity of the Internet today enables individuals to access information on SNS at any time, without temporal and geographical barriers (Bruguera et al., 2019), and to participate in topical debates (Neubaum & Krämer, 2017; Winter & Neubaum, 2016). While SNS provide a plethora of hidden data (Gundecha & Liu, 2012) and represent a tool for communication that (a) enables rapid dissemination of information (Zeitzoff, 2017) and (b) forms an online social network structure for the exchange of opinions (Bakshy et al., 2012), it is still unclear what factors influence the emergence and change of homogeneity. Previous studies have already shown that opinion leaders play an essential part in the dissemination of political information in the network (Dubois & Gaffney, 2014; Walter & Brüggemann, 2018), but they do not provide any reference to homogeneity in the network. In terms of the study of homogeneous spaces, Garimella et al. (2018) argue that a deeper understanding of this might be gained if future work focused on the interplay of network structures and the content from interactions in social media. For this reason, in addition to the influential people in the network, one should also not neglect the community structures, which in previous studies have likewise had an influence on opinion formation.

This dissertation addresses the prevalence of opinion-based and informational homogeneity in online social networks and the emergence of homogeneity from a computational social science

(CSS) perspective and presents seven empirical studies. In particular, the research field of CSS plays an increasingly important role in the study of social phenomena based on huge amounts of data (Lazer et al., 2020) and in advancing new inference based on information processing in an interdisciplinary research field (J. Zhang et al., 2020). Furthermore, CSS is also characterized by the fact that it is embedded in social science, meaning that its theories also serve a guiding function here. According to Lazar, CSS deals with human behavioral data, which uses computational methods, analyzing different types of data such as networks, images, or texts to gain new insights (Lazer et al., 2020). Methodological approaches such as SNA, agent-based modeling (ABM), and machine learning (ML) are required in this context to better visualize data, examine their network structure in more detail, model simulations based on theoretical foundations, or make predictions based on artificial intelligence (Cioffi-Revilla, 2010; Radford & Joseph, 2020). This "cocktail" of different methodological tools used in CSS, therefore, provides a suitable basis for the study of homogeneous spaces and their influencing factors in the present dissertation.

1.2. Research questions

The goal of the dissertation is to foster an understanding of political homogeneity by providing novel perspectives and methods. To pursue this goal, computational methods from the field of CSS are applied in combination with social science and communication science theories. The applied computational methods are used to investigate homogeneity on two different dimensions: First, for the investigation of opinion-based homogeneity, in which the purpose is to examine like-minded political opinions and stances in the network in light of their interconnectedness with each other. The second dimension focuses on informational homogeneity, where the focus is on like-minded information in the network and how it is interconnected. Therefore, the investigation of opinions and information disseminated by individuals in social networks on certain topics, as well as what position these individuals hold on the topic, can provide a stronger indication for determining political homogeneity, since these, as opposed to the assumption of ideological followership, contain an actual statement of how this individual thinks about the issue at a specific time. Furthermore, this can determine an accurate date and contain statements about how homogeneous or heterogeneous a debate currently taking place in the network is on a certain topic. Taking into consideration previous research, this dissertation aims to contribute to scholarship on the prevalence of homogeneity of opinions and information on various political issues and how they compare. Therefore, the

first research question is as follows:

RQ1: How high is the prevalence of opinion-based homogeneity and informational homogeneity?

Addressing this research question provides a deeper understanding of the extent to which opinions on general political topics are discussed and how homogeneously/heterogeneously they are communicated. Furthermore, the research question addresses the investigation of marginalized groups as well as fringe groups such as extreme ideological groups or conspiracy theorists, since in-group and out-group relations are taken into account to compute the homogeneity in the network. Consequently, social science theory such as the spiral of silence can also be directly adapted in this context, offering new perspectives to provide explanations for social phenomena. Secondly, answering the research question also provides space for a broader understanding of how information is disseminated in the network and how political content is proposed by algorithms.

Determining how topics are presented and discussed is the initial starting point to creating an overview of how prevalent homogeneity is in the network. However, it is also necessary to identify influencing factors that may have caused the network to exhibit strongly or weakly homogeneous communication behavior. Beyond the analysis of opinion-based and informational homogeneity, which in this way provide a snapshot of a particular political issue and platform, it is a challenge to replicate the dynamics of communication and identify potential influencing factors. Indeed, this is due to the fact that social media data is complex and makes it challenging to replicate the dynamics of the network and identify significant factors in the network. Using ABM, a virtual experiment considering multiple parameters can be built that makes it possible to capture dynamic communication patterns and their interactions in order to identify relevant factors that are responsible for how the homogeneity in the network is created and might further evolve. The use of ABM also allows for the consideration of a theory-driven implementation of social science/communication theories to build a more realistic model—hence the second research question:

RQ2: What factors are relevant for the spread of opinions and information in networks to understand the emergence and change of homogeneity?

Addressing this research question provides the opportunity to construct a scenario that mimics reality and considers the temporal aspect, in which agents with theory-based parameters are present and defined in a graph-based environment that shares the same characteristics as real online social networks, allowing for more realistic comparability and therefore enabling general conclusions to be drawn. The investigation of influencing factors, such as community structures or opinion leaders, can provide insight into how dynamic contexts of opinions emerge and change in the network, thereby allowing a greater understanding of statements about homogeneity and prevalence. Opinion leaders who take on a discrediting role and spread discrediting opinions in networks thus have a different intention and impact on how the opinion climate evolves than do ambivalent opinion leaders, where more balanced views are expressed. The dissertation examines the concept of opinion leaders as well as the consideration of community structures according to the spiral of silence theory.

Thus, the dissertation makes two contributions: First, by showing how high the prevalence of homogeneity of opinions and information on different social topics on in discussion networks YouTube is. Second, the dissertation provides information about influencing factors that emerge and change in relation to social phenomena to assess further information about homogeneity. The following section deals with the structure of the dissertation and the list of publications used to answer the research questions.

1.3. Dissertation structure and list of publications

This dissertation is a synthesis of research articles published in international journals and conference proceedings, written in a cumulative form. The structure of the dissertation is organized into five chapters:

Chapter 2 deals with the theoretical foundations for the conceptualization and measurement of homogeneity and its prevalence in online networks. Here, interdisciplinary aspects of CSS, political communication, and virtual homogenous spaces are addressed. Chapter 3 deals with the research design of this dissertation. In this chapter, the individual research papers are related and linked to the research questions. The applied methods and their data are also described in this chapter. Chapter 4 refers to the summarized research results of the findings and how the composition of the results helps to answer the research questions. Chapter 5 discusses the results and relates these to previous research, as well as taking a further look at the practical and theoretical implications of the work.

Table 1 provides a list of published scientific publications, indicating title, authors, year of publication, publication channel, and article type, with the order of articles following the logical structure of this dissertation. Other ranking factors such as VHB, SJR 2020 Ranking, and Google Scholar Citation have also been included in the table.

The journal articles are published in Computational Communication Research (CCR), Information Systems (IS), Online Social Media and Networks (OSNEM), the Journal of Computational Social Science (JCSO, accepted), and the Journal of Business and Economics (JBEC). Several of the journal articles (No. 1, No. 2, No. 3, No. 6) were also presented at the International Communication Association, which is "*the largest international scholarly network in communication.*"[1] Furthermore, the two conference papers are full papers, published as proceedings and presented at the Multidisciplinary International Symposium on Disinformation in Open Online Media (MISDOOM2020) and the SMSociety'20: International Conference on Social Media and Society (SMS).

Each article highlights aspects of various perspectives on capturing, measuring, and evaluating homogeneity in diverse spheres of communication research in combination with computational methods. These academic articles were written in English in collaboration with scholars from the University of Duisburg-Essen and the University of Edinburgh.

Table 1. List of research articles

#		Publication	Type	VHB	SJR	Citations
1	Title: Authors: Year: Publication channel:	Identifying Political Sentiments on YouTube: A Systematic Comparison Regarding the Accuracy of Recurrent Neural Network and Machine Learning Models **Röchert, Daniel**; Neubaum, German; Stieglitz, Stefan 2020 Multidisciplinary International Symposium on Disinformation	CNF	C	0.25	2

[1] https://www.icahdq.org/blogpost/1523657/285936/President-s-Message-ICA--Fair-Use?tag=October+2017

		in Open Online Media (MISDOOM)				
2	Title: Authors: Year: Publication channel:	Opinion-based Homogeneity on YouTube: Combining Sentiment and Social Network Analysis **Röchert, Daniel**; Neubaum, German; Ross, Björn; Brachten, Florian; Stieglitz, Stefan 2020 Computational Communication Research (CCR)	JNL	N/A	N/A	10
3	Title: Authors: Year: Publication channel:	Caught in a Networked Collusion? Homogeneity in Conspiracy-Related Discussion Networks on YouTube **Röchert, Daniel**; Neubaum, German; Ross, Björn; Stieglitz, Stefan 2022 Information Systems (IS)	JNL	B	0.55	5
4	Title: Authors: Year: Publication channel:	The Networked Context of COVID-19 Misinformation: Informational Homogeneity on YouTube at the Beginning of the Pandemic **Röchert, Daniel**; Shahi, Gautam; Neubaum, German; Ross, Björn; Stieglitz, Stefan 2021 Online Social Networks and Media (OSNEM)	JNL	N/A	0.65	3

5	Title: Authors: Year: Publication channel:	The homogeneity of right-wing populist and radical content in YouTube recommendations **Röchert, Daniel**; Weitzel, Muriel; Ross, Björn 2020 SMSociety'20: International Conference on Social Media and Society (SMS)	CNF	N/A	N/A	9
6	Title: Authors: Year: Publication channel:	Two sides of the same leader: An agent-based model to analyze the effect of ambivalent opinion leaders in social networks **Röchert, Daniel**; Cargnino, Manuel; Neubaum, German 2022 Journal of Computational Social Science (JCSO)	JNL	N/A	N/A	0
7	Title: Authors: Year: Publication channel:	The Influence of Community Structure on Opinion Expression: An Agent-Based Model Cabrera, Benjamin; Ross, Björn; **Röchert, Daniel**; Brünker, Felix; Stieglitz, Stefan 2021 Journal of Business Economics (JBEC)	JNL	B	0.74	2

2. Research background

2.1. Political communication in social media

In the digital society, the various social platforms play a significant role in communication between people. It is also evident that younger people are greater users of social media (Newman et al., 2021). A recent examination of multiple studies through a meta-analysis including more than 50 countries and published between 1995 and 2016 found positive evidence between digital media use and political participation (Boulianne, 2020). The comparison over time shows a continuous positive trend since 2003, which the authors explain "*by the rise of social networking sites, more interactive websites, and the rise of online tools to facilitate political participation*" (Boulianne, 2020, pp. 948-949). In particular, the exchange of opinions on social networks can motivate people to form new groups to discuss and share information on political issues (Conroy et al., 2012; Vaccari et al., 2015; Weeks et al., 2015). Likewise, social media is one of the most important news sources, alongside television (Gottfried & Shearer, 2017), and a component for the flow of political information, whereby political knowledge can be achieved through the use of SNS (Bode, 2016; Gil de Zúñiga et al., 2012). In particular, online SNS are accused in the public domain of creating virtual spaces (so-called echo chambers) where people join homogeneous communities in which they are surrounded by like-minded people who reinforce their own stance in the form of ideology or opinions and might foster processes of polarization (Boutyline & Willer, 2017; Colleoni et al., 2014; Sunstein, 2017). The assumption here is that homogeneity is particularly pronounced in marginalized groups when extreme views are involved, since studies have found that users in online networks with extreme political ideologies are more likely to be attracted to homogeneous online environments (Bright, 2016; Dvir-Gvirsman, 2017). In comparison to traditional media, social media offer new opportunities for individuals to participate in political debates within the platforms and provide fringe groups on YouTube a landscape in which to voice their opinions globally (McNair, 2017). To gain a deeper understanding of the communication of marginalized and fringe groups, the spiral of silence theory, which states that people are more inclined to express their opinions when they feel they are part of the prevailing opinion climate (Noelle-Neumann, 1974), may help to gain new insights into the research of virtual homogenous spaces. Evidence so far on the theory indicates that it can be confirmed in both offline (face-to-face) and online settings in the realm of social media (Hampton et al., 2014). Therefore, it is possible that individuals who belong to fringe or marginalized groups

and who believe, for example, conspiracy theory or extremist opinions are more likely to communicate their opinions in an online environment if they also encounter consent. In addition, YouTube videos that already disseminate such opinions could be "fertile ground" for individuals to feel encouraged in these attitudes and opinions and to participate in the debate on them, while dissenting opinions tend to be voiced less frequently.

Furthermore, Bruns suggests that social platform algorithms can also reinforce ideology, speaking of the "*algorithmic spiral of silence and reinforcement*," in which posts are displayed preferentially based on identified user patterns (Bruns, 2019a). In particular, SNS such as YouTube, which suggest further content based on algorithms according to users' behavior on the site, showed in one study that the messages suggested by the algorithm contributed to an increase in political participation (Feezell et al., 2021). This should be viewed with caution, however, as filtering gives users a restricted information environment (Bozdag & van den Hoven, 2015), which, according to the metaphor of the filter bubble (Pariser, 2012), means that algorithms are more likely to suggest personalized and thus more homogeneous content to people that is consistent with their own opinions and values than cross-cutting opinions. The existence of filter bubbles has been the subject of intense debate within the scientific community, as personalized communication should be viewed with caution and might pose a threat to society (Zuiderveen Borgesius et al., 2016). Current research shows mixed conclusions about the existence of filter bubbles. While some scholars conclude that there are no empirical findings to support the filter bubble theory on social media (Bruns, 2019b; Haim et al., 2018; Stark et al., 2020), other studies show that the findings prove the existence of the filter bubble metaphor, especially for topics of political radicalization (Bryant, 2020; O'Callaghan et al., 2015) and misinformation (Hussein et al., 2020).

In particular, misinformation has in part shaped the current COVID-19 pandemic, as users have repeatedly spread false facts on SNS about origins, prevention, diagnosis, and vaccination (Brennen et al., 2020). One reason why users in particular believe in misinformation on SNS and spread it on the network, the researchers found, is that people do not explicitly pay attention to the accuracy of the content and are more likely to act intuitively, as they may have less knowledge about the scientific evidence (Pennycook et al., 2020). In general, misinformation can be defined as: "*misleading information that is created and spread, regardless of whether there is intent to deceive*" (Treen et al., 2020, p. 2). There is rising concern among researchers as misinformation spreads faster through the network than normal information (Shahi et al.,

2020; Vosoughi et al., 2018), which could have negative consequences for society and democracy (Allcott et al., 2019; Lewandowsky et al., 2017). However, there is no empirical evidence as yet on the influence of recommendation systems in the context of misinformation and whether the accumulation of video recommendations thus creates a "chain of misinformation" in which the user increasingly receives misleading information. Although the official YouTube blog reports that this kind of “borderline content“ is being taken down (Goodrow, 2021), the recommendation algorithms and the way they operate still remain a black box.

However, previous research has shown that strategic approaches exist to reduce misinformation by spreading counter-messages, thereby correcting it with accurate information (Bode & Vraga, 2015; Chan et al., 2017; van der Meer & Jin, 2020). Nevertheless, the attempt to counter misinformation can also lead to backfire effects, where the original beliefs are reaffirmed and reinforced (Lewandowsky et al., 2012). One approach that may be of interest in combating misinformation is so-called information-rich actors, such as opinion leaders (i.e., people who have a strong influence on public opinion, such as journalists and politicians), who can reduce mistrust and ease sense-making (Mirbabaie et al., 2020). Detecting opinion leaders in social networks is challenging and has been investigated in many studies using a variety of approaches (Borge Bravo & Esteve Del Valle, 2017; Oueslati et al., 2021; B. Zhang et al., 2020). According to Lazarsfeld et al. (1944), opinion leaders are characterized by having a higher probability of influencing the political opinion of their social network since they have greater political interest, which can reflect an important aspect for opinion formation in social network groups. Furthermore, they have a key role in online discussion landscapes as they are pivotal in the dissemination of opinions (Himelboim et al., 2009) and, therefore, can affect the credibility of information and their social contacts within the network in interpersonal communication (Turcotte et al., 2015). From a network technology perspective, opinion leaders in social networks are characterized by their high degree of connectedness to other nodes (Nazir et al., 2008; Wattenhofer et al., 2012). In this context, however, the question arises as to what role opinion leaders generally play and what the spectrum of opinion looks like in this respect. In most cases, opinions are often divided in a one-dimensional way into "pro" and "contra," which in a nutshell does not reflect reality. However, the concept of "attitudinal ambivalence" states that attitudes are represented in a two-dimensional space, in which positive and negative attitudes are taken into account, but are independent of each other in their own dimension (Armitage, 2003; Schneider & Schwarz, 2017; Thompson et al., 1995). To calculate the

ambivalence of the two dimensions, an averaging of the positive and negative attitudes is applied (Huckfeldt et al., 2004). Thus, opinion leaders who have an ambivalent attitude (e.g., public broadcasters) could ensure that people in the network are influenced with opinions in a more balanced way since they receive, for example, arguments "in favor of a political topic" as well as "counter-arguments," so that people are able to form their own opinions based on the given arguments. On the other hand, it would also be possible that there are discrediting opinion leaders in the network, who discredit the viewpoint of other opinions and spread misinformation in order to gain an advantage and win the opinion climate for themselves. Results on the investigation of the 2016 U.S. presidential election on Twitter found that opinion leaders spreading misinformation and distorted news were characterized by accounts that were not verified and had false profiles (Bovet & Makse, 2019), which can also make identification even more challenging. Likewise, the findings indicated that this political misinformation originated from a networked cluster and was spread by several small groups as a collective. This suggests that it is not only the identification of influential nodes that is important, but also how nodes are connected within a network and potentially form communities. For this reason, it is necessary to investigate communities and influencing factors such as opinion leaders and the attitudes (ambivalent, discrediting) they hold in the network, as these may explain relationships to the emergence and change of homogeneity.

2.2. Homogeneous information spaces

The formation of virtual homogeneous spaces is often explained by the social phenomenon of selective exposure (Colleoni et al., 2014; Knobloch-Westerwick, 2014; Zillmann & Bryant, 1985) and associated with the consequences of political polarization (Bakshy et al., 2015; Conover et al., 2011). Previous research examining political homogeneity on Facebook and Twitter has demonstrated that people in the U.S. are connected to a higher degree with like-minded users than with politically opposed users (Bond & Messing, 2015; Boutyline & Willer, 2017). Despite this high degree of connectedness, another study that focused on communication and interaction data from Facebook users and their friendship network highlighted that 20% of U.S. Facebook users are affiliated with opposing parties and receive content that does not align with their ideological beliefs (Bakshy et al., 2015). The investigation of these potential homogenous spaces has so far been pursued by two research approaches, one using quantitative analyses in the form of questionnaires to ask social media users about their behaviors and perceptions, and the other using content-oriented analyses, in which data from social media are analyzed using computational methods. Here, quantitative research comes to the general

conclusion that people on social media are occasionally exposed to heterogeneous opinions (Kim, 2018; Lee et al., 2014; Lu & Lee, 2019; Vaccari et al., 2016). In this regard, the study by Geiß et al. (2021) demonstrates, first, that social media does not have a reinforcing effect on the expression of opinions, but only on the use of political information and, second, that only certain groups holding extreme opinions have a higher tendency to drift into homogeneous spaces. Furthermore, the researchers found that in particular people who express extreme attitudes are more motivated to express their opinion if (a) many political actors and parties are represented in the network and (b) a high level of engagement in social media is associated with their opinion. These results are in line with previous literature (Bruns, 2019a, 2021; Dubois & Blank, 2018).

However, there is also literature showing contradictory findings about the existence of virtual homogeneous spaces. A study by An et al. (2019) used the linguistic characteristics of users in relation to political discussions on Reddit. The researchers wanted to know, based on Reddit interactions and communication spaces, to what extent users encounter different partisan political attitudes and whether the emergence of homogeneous spaces is possible. The findings revealed that supporters of opposing candidates (Hillary Clinton and Donald Trump) are not engaged in homogenous discussion cocoons, but are characterized by heterogeneous cross-interaction communication. However, the findings also demonstrated that only a minority of users on the Reddit platform who supported Hillary Clinton or Donald Trump were active in politically homogeneous communications, which, as the authors stated, can co-exist on Reddit (An et al., 2019). In content-oriented research, however, there is only weak evidence indicating the existence of virtual homogenous spaces on Twitter (Bakshy et al., 2015; Boutyline & Willer, 2017; Colleoni et al., 2014). A recent study focusing on a multi-platform analysis to investigate homophilic clusters in social online dynamics demonstrated that the emergence of homogeneous spaces is favored when "*platforms organized around social networks and news feed algorithms*" (Cinelli et al., 2021, p. 2). This raises the question, however, of the extent to which user preferences about their political views can be realistically classified based on likes of posts or mentions of links in news outlets. Therefore, further efforts are needed to collect data from social networks and analyze these with new network technologies in order to understand how people communicate within social networks, as well as their structures across the Internet, and what dynamic relationships and interactions can emerge from them (Garton et al., 1997; Hogan et al., 2008). Comparing the literature, however, it is evident that there is no unified procedure for the interpretation and methodological analysis of homogeneous spaces.

2.2.1. Opinion-based homogeneity

Previous research focusing on the study of homogeneity in political virtual spaces has so far been very well documented in terms of empirical results, and these results are mainly directed at ideological homogeneity, where the focus is on moral values and political identities (Bakshy et al., 2015; Barberá et al., 2015; Del Valle & Bravo, 2018). Ideological homogeneity seems to be useful for the consideration of a two-party system (Democrats, Republicans) and the views of members within the network, and could already be measured on online social platforms with network analysis. However, the question arises as to how accurate these results are based on Twitter networks when followers have the same ideological viewpoint as their politicians, since the actual ideology of users is barely discernible in online SNS. On the other hand, there is a vast amount of textual information on diverse topics on SNS, which allows users to express their opinions by means of comments and thus, if necessary, form connections with other users and comments (Lange, 2007). Thereby, people may already be members of ideological groups (parties) from which the opinion may have formed and want to express this opinion in textual or linguistic form (Dijk, 1995).

In addition to the political contexts and groups that have been mentioned so far, there have also been studies addressing the issue of the prevalence of conspiracy theories on social media. Previous studies in this respect assume that a homogeneous communication environment might also emerge between conspiracy believers (Garrett & Weeks, 2017; N. Smith & Graham, 2019). In studies examining Facebook and YouTube data, it seems that the use of conspiracy-theory content could be shown to be associated with the polarization of users and their presence in homogeneous communication networks (Bessi et al., 2015, 2016; Del Vicario et al., 2016). However, this is not consistent with recent findings from a study of Facebook users, which found no evidence of a link between online network homogeneity, populist attitudes, and conspiratorial beliefs (Cargnino, 2020). These fringe groups become important for the prevalence of homogeneity since they reflect a part of society that de facto represents a minority opinion. Attitudes in the form of comments and videos advocating conspiracy theories can thus also have an influence on other users, as it is suggested that the content serves as a baseline for society as a whole (Neubaum & Krämer, 2017). Therefore, user-generated content (UGC) needs to be considered for homogeneity, as it contains opinions and attitudes that are directly related to the topic. In research, the potential of textual data in the form of opinions and information from social media is well known; however, challenges still remain in this regard (Stieglitz et al., 2018), and doubts exist regarding analysis for homogeneity: "*even though collecting opinion*

data is expected to provide more advantages than relying on demographic proxies, the method is not applicable in the case of scientific publishing" (Vuong et al., 2021, p. 3). This dissertation therefore addresses the concept of opinion-based homogeneity for the study of homogeneous spaces in social networks, which allows us to relate the opinions and attitudes of users to political and social issues and the structures by which they are communicated, thereby enabling a consideration of the dynamics and processes of discussions. Connecting similar information snippets in the network allows the computation of in-group and out-group relations, which provides a reliable indication of the homogeneity in the network. Here, it is also possible to include not only comments in the analysis, but also videos that reflect a higher-level opinion on a particular topic. Thus, opinion-based homogeneity makes it possible to include the interrelationships of semantic content that are interconnected within the discussion network in order to make statements about homogeneity and heterogeneity based on specific political topics.

2.2.2. Informational homogeneity

An SNS is not only a space to exchange opinions, but also a space where users obtain information about social and political issues and share this with other users (Ellison & Boyd, 2013). In addition to the numerous serious sources of news offering users good quality information, many dubious sources still exist that deliberately disseminate misinformation. Even before the COVID-19 pandemic, the problem of misinformation in social media on political (Badawy et al., 2018; Kušen & Strembeck, 2018) or health information [e.g., vaccination (Donzelli et al., 2018)] existed. Throughout the course of the COVID-19 pandemic, however, the extent of misinformation found on a wide variety of social platforms became more apparent (Cinelli et al., 2020). News articles also reported on the dangers of homogenous spaces on YouTube in light of the current COVID-19 pandemic and how they are related to the cross-cutting movement and the spread of misinformation that can potentially radicalize users. Although the findings of the studies indicate the existence of misinformation, communication between individual users is not considered, with the result that there is currently no evidence on the homogeneity of misinformation. In this regard, studies have demonstrated that regular use of social media can lead to a greater persuasive effect for misinformation, as was found in relation to COVID-19 (Allington et al., 2020; Su, 2021). Hence, one needs to introduce the concept of informational homogeneity, which addresses the degree to which similar types of information are interconnected. The connectedness of users spreading misinformation can be

directly linked to other misinformation as well as non-misinformation in the communication network, taking into account in-group and out-group circumstances.

While the existence of misinformation is a problem, there is some concern among researchers that personalized algorithms may also pose a threat to society; the reason for this is the assumption that personalized content can reduce the diversity of online content (Zuiderveen Borgesius et al., 2016) and restrict users' ability to independently choose information (Bozdag & van den Hoven, 2015). In the worst case, this can lead to users tending toward extreme ideologies or reinforcing them (Sunstein, 2007). According to a recent study by Kaiser & Rauchfleisch (2020), YouTube's recommendation algorithm may encourage the formation of highly homophilic communities and lead users into far-right spaces. However, some researchers also argue that the YouTube platform is less likely to provide people with similar and homogeneous political information since it is a heterogeneous landscape of viewers who are exposed to a variety of messages (Evans, 2016). Having said that, the results of this study rely only on YouTube "likes" and "dislikes" as indicators, which serve as a feedback mechanism for viewers' attitudes. However, the attitudes of users in text form, such as within comments, is not taken into consideration here. Aspects of how homogeneous the proposed contents of the recommendation system might be had not been considered in previous studies. The principle of informational homogeneity can also be adapted to other use cases to analyze the behavior of algorithmic recommendations with respect to their homogeneity. In the case of recommendation systems, knowledge of how the algorithms behave in terms of using their filtering functions might be improved in order to increase understanding of their implications, given that users have limited understanding of this (Bucher, 2017; Powers, 2017). In the following section, the approaches used as baselines for the computation of the two presented homogeneity calculations are explained.

2.2.3. Computation of homogeneity

The aim of homogeneity computation is to understand how people communicate within social networks, as well as the SNS structures across the Internet, and which dynamic relationships and interactions can emerge from them (Garton et al., 1997; Hogan et al., 2008). For the study of homogeneous spaces in social networks, Bruns (2021) argues that it is necessary to place the methodological focus on the network structure, since in this way, the interaction and communication of users is based on the comparison of in-group and out-group connections, which can be achieved by calculating the external–internal (E-I) index (Krackhardt & Stern,

1988). In doing so, these in-group and out-group connections can be applied to different topics (e.g., climate change), since nodes in the network reflect users with opinions (e.g., pro, contra) interacting with each other in the network, and it is also possible to make statements about the homogeneity of communication. To identify the opinion, content analysis is needed, which is "*a research technique for making replicable and valid inferences from data to their context*" and annotated by trained human coders (Krippendorff, 2013, p. 24). In particular, the annotation of the data is an important step within natural language processing (NLP), in order that the information can be learned by the computer (Pustejovsky & Stubbs, 2013). The computation of the E-I index can be calculated for the entire network and across all classes (global E-I index) or isolated for the individual classes in the network (class E-I index). Using the permutation test, a statistical significance test can be performed to test the null hypothesis on the observed data set, whereby many random permutations of the network are computed to calculate the statistical significance of the expected data set (see. Scott & Carrington, 2011). The methodological approach specified here for computing homogeneity assumes, that data were collected from SNS, taking into account that data collection is dependent on the platform's own application programming interface (API), which has its own limitations (Geissinger et al., 2020; Stieglitz et al., 2014). ML/DL methods are also essential, offering the ability to build models that can automatically improve themselves based on Big Data (Jordan & Mitchell, 2015). However, previous studies that have looked at homogeneity in Twitter networks and used the E-I index as a benchmark have reached different conclusions. Del Valle & Bravo (2018) analyzed the Catalan parliamentary Twitter network and found a tendency toward party political and ideological homophily, whereby delegates are more willing to interact with party members holding the same political interests. In contrast to Del Valle & Bravo, Bruns (2017) analyzed 255,000 Australian Twitter accounts and showed that there was only a moderate level of homogenous follower/followee connections through clusters, therefore not indicating a strongly homogenous space.

The studies presented here provide a first indication of how homogeneity in social networks can be calculated. However, comparability is limited since different terminologies are used to examine virtual homogeneous spaces in this field and therefore no uniform understanding exists. Likewise, the comparability of the presented studies is somewhat problematic, as they mainly deal with the ideological relation to each other and thus consider less the content aspect of opinions and information. Thus, to understand the prevalence of homogeneity of opinions and information in social networks, and likewise, to understand what factors influence

homogeneity over time, it is necessary to develop new methodological approaches. Accordingly, this dissertation develops new methodological methods in light of social science and communication science.

2.3. The role of computational methods

The digitization of the world and the urge to interconnect the globalized world have generated not only a new social interest, but also the term "Big Data" to create new possibilities regarding social concepts and decision-making mechanisms (Bachmann et al., 2014). In view of the growing flood of data, the emerging research field of CSS is gaining relevance to investigate and analyze the communication of digital media and its data with an interdisciplinary repertoire of computational methods (Edelmann et al., 2020; van Atteveldt & Peng, 2018). Here, the goal is to examine vast amounts of human communication data from social media, such as comments or videos, using these computational methods in order to a) gain a better understanding of social phenomena, b) develop new theories, and c) better understand human behaviors and social dynamics in relation to different social and political issues (Edelmann et al., 2020; Lazer et al., 2020; J. Zhang et al., 2020). Researchers are also already drawing links to new subfields of CSS that focus in particular on political communication (Theocharis & Jungherr, 2021) in order to consider issues and problems associated with digital media from different perspectives and angles. Researchers claim that the study of social media requires the use and development of new computational methods capable of analyzing vast amounts of data to gain deeper insight into interactions within networks (Bruns et al., 2011).

Computational methods can be pivotal in examining current concerns regarding the use of digital media for information diets, racism, and xenophobia, as well as misinformation using citizens' communication data, to gain further insights into these extreme scenarios in research (Theocharis & Jungherr, 2021). For these insights to become apparent, methods used in social networks might be applied to map human relationships and their communication paths to each other (Chen et al., 2014) in order to determine, for instance, how information spreads within network structures in user communities (Croitoru et al., 2015) or what topics are associated with communities (Reihanian et al., 2016). SNS not only have a significant impact on the dissemination of information and opinions, but also, according to Bakshy, can dramatically change users' attitudes when they come into contact with new information (Bakshy et al., 2012). A core aspect of this is that information is analyzed more precisely according to its content in order to generate knowledge from communication data (Krippendorff, 2018). User-generated

data, for instance, can offer valuable research opportunities by reflecting content and interactions that can be used to investigate different aspects of research, such as political homogenization and polarization. Furthermore, methods of text analytics can be used to transform unstructured data into computer-readable information and further advanced using ML approaches to recognize and automatically process meaning relationships in texts. Previous studies have already shown successful implementation of hybrid computational methods, which were analyzed in the field of social media using text analysis and network analysis on the topics of eating disorders and crisis management (Moessner et al., 2018; Romascanu et al., 2020). Lewis et al. (2013) argue that a combination and thus a hybrid approach of different computational methods and manual annotation can lead to an improvement of traditional content analysis.

In addition to the techniques that examine the content of users in existing SNS, there are also other CSS techniques, such as ABM. In ABM, interactions between agents are modeled on the basis of predefined parameters and rules to investigate dynamic processes and thus to understand "*how individuals and the environmental variables influencing them vary over space, time or other dimensions*" (Railsback, 2019, p. 11). Therefore, agent-based models can be found in a wide range of research disciplines, as they offer versatile applications and generate better knowledge of how complex systems will behave in the long term and what effects they are associated with (Wilensky & Rand, 2015). Likewise, ABM can help study dynamic micro and macro processes (Bruch & Atwell, 2015; Waldherr & Wettstein, 2019), e.g., in the context of opinion formation. Studies that have used ABM have demonstrated that it is feasible to model currently emerging communication scenarios, such as the presence of social bots in social media (Ross et al., 2019), and thus measure their impact on opinion formation (Sohn, 2019). In general, computational methods offer a huge repertoire of different analysis possibilities to study human behavior and communication, given the ever-increasing amounts of data. Leveraging Big Data analytics thus indicates the potential to investigate correlations and patterns of human behavior on a large scale, which would be challenging using laboratory experiments as well as time resource-intensive (Qiu et al., 2018). This combination is by no means intended to replace existing research in social or communication studies, but rather to enhance their perspective and demonstrate that social phenomena can be viewed and analyzed from a variety of perspectives.

This dissertation addresses this aspect and provides not only the main content-related added value that has been explained in the previous two sections, but also methodological added value. This methodological added value is characterized by the fact that an interplay between data-driven and theory-driven approaches is adhered to, in which a variety of CSS procedures in conjunction with Big Data datasets were used to identify patterns and correlations of human online communication and to provide a sufficient explanation of homogeneity in networks.

3. Research design

This chapter is divided into two main sections that cover the research strategy and the applied research methods. The first section starts with the conceptual research strategy that was used for the cumulative work and provides an overview of the way in which the individual research articles are linked to each other and how they relate to the research questions. The second section explains the research methods used in terms of data collection and data analysis.

3.1. Research strategy

Figure 1 below shows the composition of the research articles to answer the research questions in this dissertation. Paper P1 provides a foundation for the other papers (P2–P4), as it elaborates methodological approaches. Here, YouTube, as an online social platform, is used to investigate the computation of homogeneity of opinions (P2, P3) and information (P4, P5) within videos and comments, as well as their relationship to each other. Therefore, statements about homogeneity and its level of prevalence have been made in the articles on politically controversial topics (P2), conspiracy theory content (P3), misinformation about the COVID-19 pandemic (P4), and radical right-wing content (P5). The papers also discuss the explicit distinction between opinion-based homogeneity and informational homogeneity. The last two papers (P6, P7) address the second research question, which deals with influencing factors that are relevant for the spread of opinions and information in networks and which emerge and change homogeneity. Using ABM, social phenomena such as opinion leaders (P6) and the spiral of silence within communities (P7) could be considered in the analysis to identify influencing factors that can affect the opinion climate.

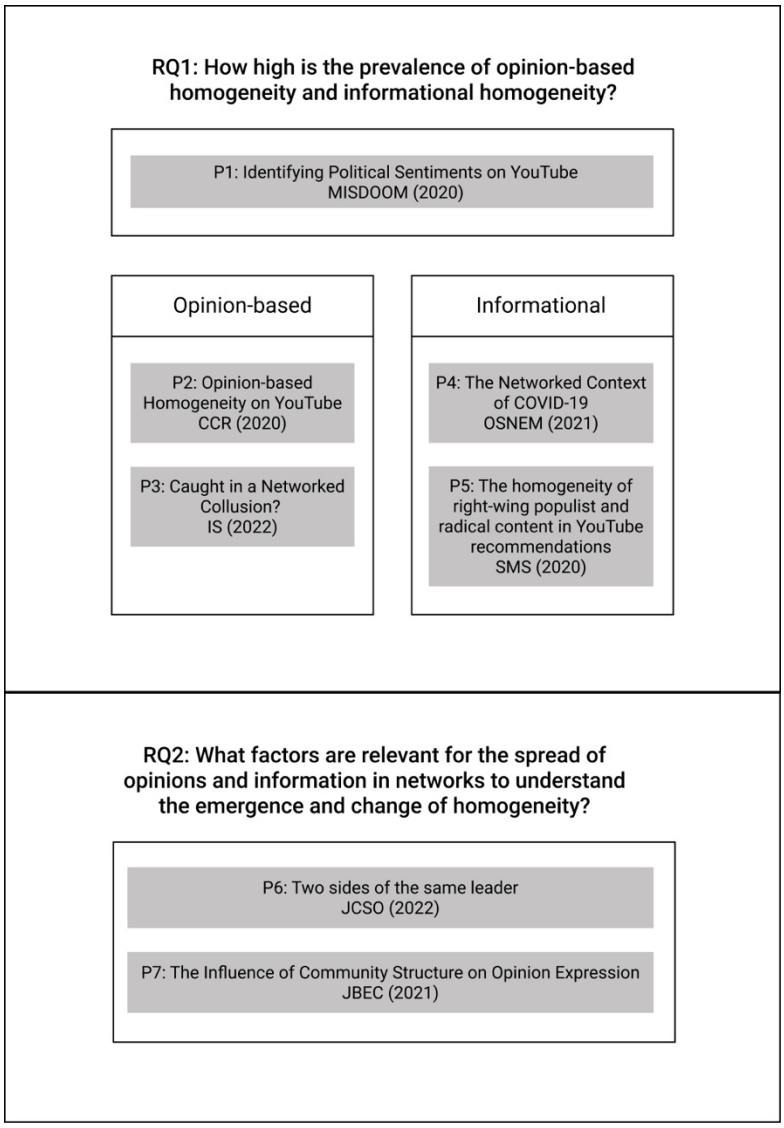

Figure 1. Overview of research questions and the related research articles

3.2. Applied research methods

Table 2 below provides an overview of the applied methodological procedures that were required to address the research questions of this dissertation. It references the individual papers with their research approach, their data source, as well as the analysis method used. It is notable that to answer the first research question, all data sources originate from the YouTube API. Although these data covered diverse subjects, it is nevertheless relevant in this case to clarify the rationale for their inclusion. For this reason, I would like to address the relevance of the platform and explain why the research focus in this dissertation is on the YouTube platform. First, it should be mentioned that previous research indicates that studies in the field of political communication is mainly focused on traditional media or online social platforms such as Twitter, while other platforms in this context have been studied to only a limited extent (Van

Aelst et al., 2017). For this reason, it is even more important to examine alternative platforms in order to observe political communication between users there, as well as to compare this with other platforms. The second reason for studying the YouTube platform is the interaction dynamics of users, who are in a content-driven relationship (Wattenhofer et al., 2012), and thus the communication is focused on the topic in the videos. This topic-specific presentation in the form of visual information forms a new main component in today's social networks (Newman et al., 2021). In addition to the content-driven relationship, YouTube also addresses a strong social component, which promotes the formation of communities (Burgess & Green, 2018) and thus provides the opportunity to communicate with other users (Lange, 2007). Furthermore, the YouTube platform has become an important source for general news gathering. According to Reuters' "Digital News Report 2021," the weekly use (for any purpose) of the YouTube social network was 53% in 2014, which had increased to 62% by 2021, making it the most frequently used SNS. Here, the motivation of YouTube users is more focused on getting an alternative perspective to the ordinary mainstream news (Newman et al., 2021). With recent reports of increased radical content and misinformation on the platform, this serves as an excellent basis to also examine the homogeneity of opinions and information from fringe groups in terms of their interaction and network structures. Considering political communication on the YouTube platform, the current arguments indicate that an investigation appears to be necessary due to the content and its network structure. Since the platform is also filled with different types of data, this increases the possibilities of analysis by means of computational methods.

In addition to providing the rationale for the investigation platform, the remainder of this dissertation deals with a description of the research methods that were essential to answering the research questions. In particular, a wide range of computational methods were applied, which in many respects can be seen as mixed methods, since a sequential application of methods was necessary in some of the research articles. For example, NLP techniques were applied to annotate textual data for further processing. For the respective annotation tasks, several coders were provided with a codebook in order to meet the same requirements and ensure that everyone was on the same level of knowledge. After completion of the coding, intercoder reliability was calculated in order to evaluate their annotation and to obtain a more accurate result for the determination of the classes of the comments and videos by means of majority voting. ML techniques were used to train and evaluate models, as well as to make predictions based on the entire data set. Since not only one model was applied for the training, a performance comparison with different models [support vector machine (SVM), logistic regression (LR), recurrent

neural network (RNN), long short-term memory (LSTM), bidirectional encoder representations from transformers (BERT)] was always aimed for—thus the best model was determined. SNA was then used to examine the interactions between in-group and out-group connections and calculate their homogeneity. The computation of the E-I index was used and further modified for direct networks. This methodological combination of NLP, ML, and SNA was necessary to determine the prevalence of homogeneity of opinions and information on the YouTube platform and thus to gain a deeper insight into the contextual dimension.

As a further research method and for the investigation of the second research question in the dissertation, ABM was used in papers P6 and P7. Waldherr and colleagues point out that "*the computer simulations of ABMs are virtual laboratories that help formalize and explore dynamic, multi-level theories of communication*" (Waldherr et al., 2021, p. 248) and can help to better understand the gap between dynamic micro and macro processes (Waldherr & Wettstein, 2019). Using ABM techniques in this dissertation, the temporal development of an opinion climate within social networks was considered, in which the opinion formation was created in the content context of social phenomena (spiral of silence, opinion leaders) to thus draw conclusions about factors that are responsible for the emergence and change of homogeneity. Accordingly, for the identification of these factors, it is important to establish a dynamic environment that considers communication structures changing over time and makes it possible to determine interventions based on parameters in order that numerous scenarios can be simulated. Indeed, it can be argued that simulation studies are only simulated data that do not reflect reality, as they have not been collected through experiments or surveys; however, it can just as well be argued that agent-based models are formed from theoretical derivations in which parameters that have been identified through previous research are determined. According to scholars, the challenge arises, on the one hand, to properly represent the existing social theories (which can also be partially incomplete) in an agent-based model with the individual regulations and parameters, and, on the other, to determine a realistic and target-oriented parameter range (number of agents) for on-demand computing time and to avoid unnecessarily extensive data sets (Squazzoni et al., 2014; Waldherr & Wettstein, 2019). However, the methods of ABM offer an essential instrument to understand potential scenarios that can be varied and have a wide range of parameters, thereby representing an alternative investigation that can reveal empirical evidence compared to classical field experiments and surveys with few resources. ABMs can also advance research by generating new hypotheses and thus providing a better understanding of social phenomena (Waldherr et al., 2021).

Table 2. Overview of the applied methodological approaches

Paper	Research approach	Data collection method	Data analysis method
P1	Methodology Comparison	YouTube API	NLP, ML
P2	Social Media Analytics	YouTube API	NLP, ML, SNA
P3	Social Media Analytics	YouTube API	NLP, ML, SNA
P4	Social Media Analytics	YouTube API	NLP, ML, SNA
P5	Social Media Analytics	YouTube API	NLP, SNA
P6	Virtual experiment	Simulation	Quantitative summary, SNA
P7	Virtual experiment	Simulation	Quantitative summary, SNA

4. Research results

This chapter presents the findings of the individual research articles. The findings are presented here in sequential order, as described in Section 3.1. Tables 3 and 4 briefly summarize the core findings of each research article. Papers on RQ1 (*How high is the prevalence of opinion-based homogeneity and informational homogeneity?*) (Table 3) covered the methodological approach and their combined use of NLP, ML, and SNA to investigate the computation of homogeneity on different topics. Since Paper P1 does not directly address the research question on RQ1, but does provide preliminary methodological work for P2–P4, the results are considered separately from the research question and, accordingly, are not presented in Table 3. Paper P1 deals with the systematic methodological comparison of text classification of German-language political YouTube comments. The performance comparison based on the F1 score revealed that the use of word embeddings yielded better results for the RNNs than for the ML models.

Table 3. Summarized results of research articles addressing RQ1

Paper	Summary
P2	The authors of this paper computed the opinion-based homogeneity on the basis of German YouTube comments by using a combination of NLP, ML, and SNA methods on three controversial political topics. The findings revealed a moderate level of heterogeneous connections, indicating that a heterogeneous opinion climate existed on these topics, where users' opinions were associated with dissimilar rather than similar opinions.
P3	This paper deals with the investigation of opinion-based homogeneity of discussion networks for three conspiracy theories (Hollow Earth, Chemtrails, and New World Order) on YouTube. The results showed that people who expressed a favorable stance toward a conspiracy theory tended to respond to content from or interact with users that shared the same opinion. In contrast, users who challenged conspiracy theories interacted in more heterogeneous discussion networks (with the exception of opponents of the Chemtrails theory).
P4	During the COVID-19 pandemic, misinformation was increasingly prevalent on social media. This paper computed informational homogeneity based on the YouTube network through videos and comments from January to March 2020, which allowed the authors to determine how heterogeneous/homogeneous the

	discussions were among users. Here, the concept of "informational homogeneity" is introduced, which makes it possible to measure the degree of homogeneity to which misinformation (as opposed to non-misinformation) is directly linked to other misinformation content (i.e., comments, replies, or videos) in a network. Furthermore, not only the fragmentation in the network is considered for the individual months, but a distinction is also made between two types of networks, which consist only of the communication of comments and replies, as well as the entire networks in which the communication of videos, comments, and replies are included. In both cases, and bearing in mind the fragmentation of the network, the findings indicated that misinformation regarding COVID-19 also exists on YouTube; however, the interconnectedness among users to discuss misinformation in the network is highly heterogeneous.
P5	While the previous papers examined homogeneity to active communication in the form of comments in the context of videos, the focus here was on examining YouTube's recommendation algorithm. This paper examined the recommendation behavior of populist right-wing and politically neutral videos on the YouTube platform in order to investigate their homogeneity. The network analysis based on the YouTube recommendation network demonstrated that the probability of being recommended another right-wing populist video after watching a right-wing populist video is 54%. However, after following the recommendation, the probability of being recommended the next right-wing populist videos drops to 37%.

Papers addressing RQ2 (*What factors are relevant for the spread of opinions and information in networks to understand the emergence and change of homogeneity?)* dealt with opinion dynamics in social networks, which were simulated using ABM, and considered the effects of opinion leaders and the spiral of silence mechanism in community structures. The aim was to identify influencing factors through a temporal evolution of the opinion climate in order to evaluate their formation or transformation of the environment. While in P6 it is opinion leaders that are the key players, in P7 it is the community structure.

Table 4. Summarized results of research articles addressing RQ2

Paper	Summary
P6	In this paper, the authors conducted a virtual experiment using ABM to investigate the influence of opinion leaders on the opinion climate in social networks. It was shown that opinion leaders have an influence on the opinion climate. Opinion leader characteristics such as ambivalence (an identical number of arguments regarding the respective opinion camps is given when expressing opinions, so that both sides are equally favored) and discrediting (discrediting the opponent's position by spreading negative opinions to neighboring agents) highlighted the fact that discrediting the opposite side leads to a majority distribution of opinions and that ambivalent opinion leaders contribute to a balanced opinion climate.
P7	In the context of the spiral of silence and the influence of community structures with their connectivity, another virtual experiment was conducted in this paper using ABM. One result of this paper was that smaller, more fragmented communities lead to minority opinions prevailing in the network. The second finding demonstrated that the more interconnected the communities were, the stronger the spiral of silence effect was.

4.1. Opinion-based and informational homogeneity

This section deals with the results of the papers P1–P5, which investigated how present homogeneity in opinions and information on various politically and socially relevant topics is on YouTube. In this section, I present the results of P1, which focuses on a methodological comparison of ML models, followed by an explanation of the schematic structure of the analysis, which describes the relationships. Finally, the results of the research articles P2–P5 are presented.

Before the investigation of homogeneity in OSN started, a deeper understanding of the various ML technologies needed to be obtained to establish methodological comparability. This is necessary to evaluate the different methods and draw conclusions on how the sample of real, unstructured YouTube datasets relate to the models and their performance with respect to each other in order to select the best model. Thus, the goal of P1 was to compare the performance of ML models based on word embeddings (word2vec, fastText) and different techniques (Skip-Gram, CBOW) to identify which models perform best in correctly predicting YouTube comments. For the methodological comparison, the 22,720 comments on two controversial political topics (adoption rights for homosexual couples, wearing religious headscarves) on YouTube were examined and annotated based on three classes (positive, negative, other). Based on the two datasets, individual word embeddings were created that served as further input for the models: RNN, SVM, and LR in addition to the training data. Figure 2 below provides an overview of how the data was preprocessed, trained, and tested. However, it is important to note that DL models like BERT, which perform extremely accurately in text classification, were not included in the methodological comparison since they were not yet available at the time of implementation.

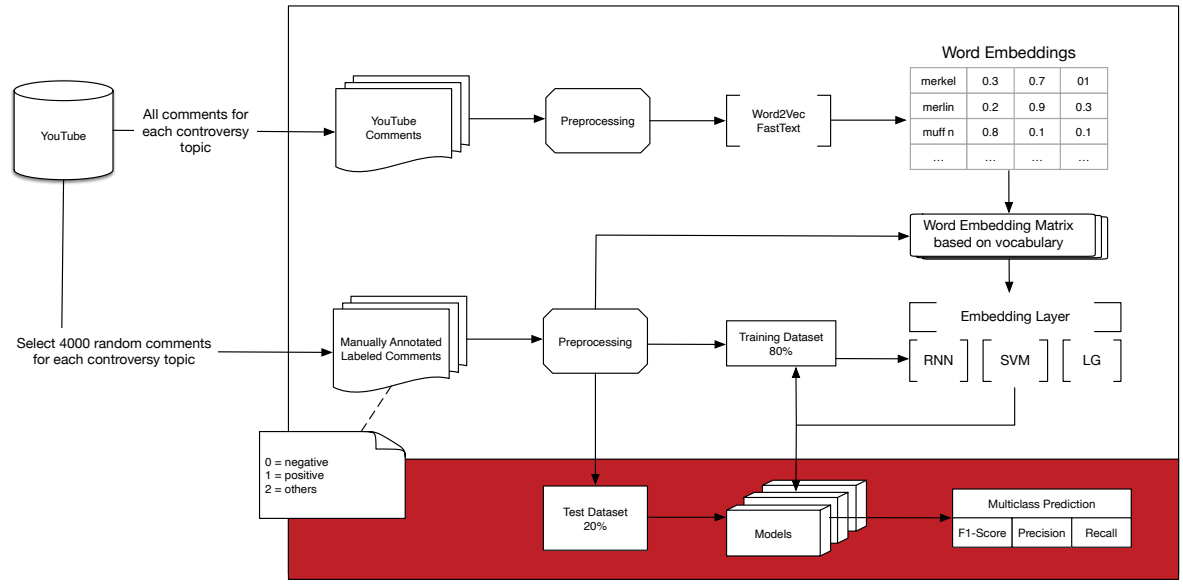

Figure 2. Workflow of the data analysis (P1)

In the methodological comparison, the authors found that based on the weighted F1 score, all RNNs outperform the traditional ML models. Furthermore, no differences in the performance of word2vec and fastText could be found, but the technique CBOW performed better than Skip-Gram, especially for RNNs. P1 argues that based on the small and imbalanced dataset, the DL models perform better than ML models, but further research would be advisable if the

implementation of the traditional ML is not based on word embeddings but on term frequency times inverse document frequency vectors (TF-IDF) to allow further comparability.

Before examining the homogeneity of opinion and information in more detail in papers P2–P5, a common understanding of the analysis is required, as this involves an interplay of different methodological procedures (see Figure 3) and is divided into three nested layers: social media platform, data, and analysis. The YouTube social media platform provides the foundation. Users of this platform visit it to obtain information in the form of videos and to engage in social interactions by expressing their opinions in the form of comments and responses on topics. This data can be collected using software that provides access to the YouTube API and contains important metadata that is needed for further analysis to compute homogeneity. Therefore, the analysis is divided into three core aspects: NLP, ML and SNA, which apply different procedures to process and subsequently analyze the data. In this context, procedures such as data annotation of samples, data preprocessing of textual data, or topic modeling are classic tasks that have been used with NLP. Training and fine-tuning of models are instead tasks that deal with ML to predict the whole data set in the further process. SNA, on the other hand, was used to transform the data into a network structure, determine its parameters, and subsequently calculate its network homogeneity using the E-I index.

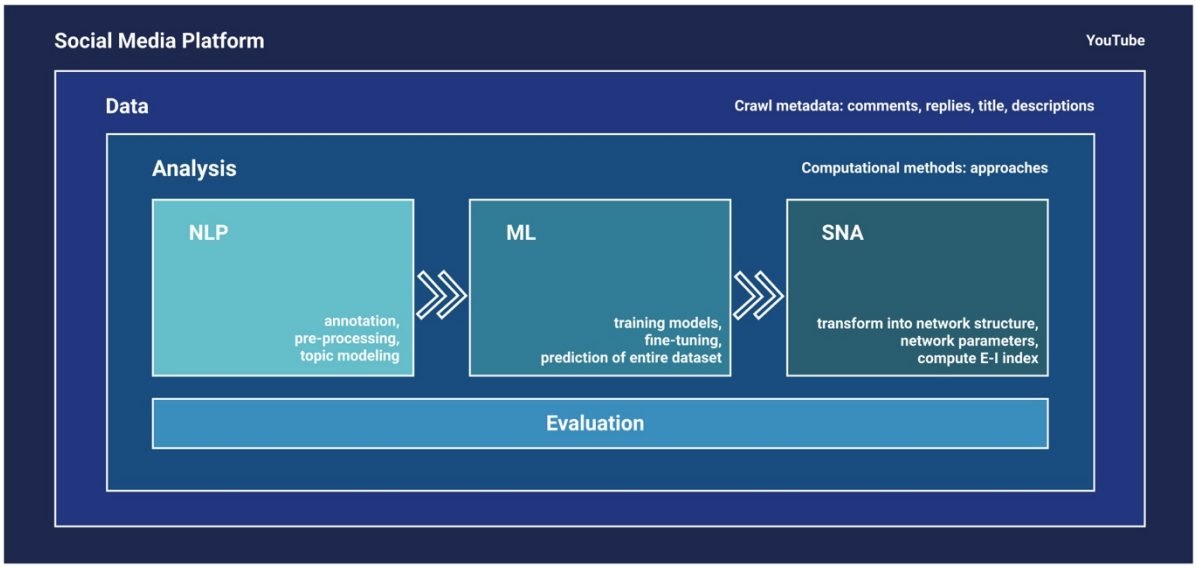

Figure 3. Schematic structure of levels to investigate homogeneity

For the computation of homogeneity (opinion-based, informational) in papers P2–P4, an ML method is used in which a model is trained on annotated comments/video information, and the entire data set is predicted using this model. This data is transformed into a network structure

and the homogeneity is then calculated using the E-I index. To better represent the communication behavior of incoming nodes and their edges, this was adapted for the calculation of the E-I Index, since the E-I Index is mainly used for undirected networks. With the modification and consideration of incoming edges, a more accurate representation is shown, which is closer to reality. The E-I index can have a value range of –1 and +1. In relation to Table 5, a value of –1 means that it is a homogeneous discussion network. Here, all connections between users (nodes) are related to their own particular class, while a value of +1 reflects a heterogeneous discussion network, where there are no connections of users (nodes) to the same class. A value of 0 indicates that all connections occur equally. The interpretation of these values in relation to homogeneity can therefore be as follows: if the E-I index values lie between 0 and –0.33, one speaks of weak homogeneity, whereas values between –0.33 and –0.66 indicate moderate homogeneity. An E-I index below –0.66 would accordingly show strongly homogeneous behavior. Conversely, the positive value intervals of the E-I index can explain the different levels of heterogeneity. As a further evaluation step, a permutation test was also performed in the studies P3–P5. To calculate informational homogeneity in P5, the ML point was skipped, since here the data set was annotated, which does not require further efforts to train and predict data.

Once the methodological approach for the classification was known, P2 introduced the concept of opinion-based homogeneity. The motivation of the paper is to provide a different perspective compared to previous research approaches that refer to ideological homogeneity in online networks (e.g., are Republicans more likely to be connected to Republicans?), since the previous methods are based only on a general group level and refer to moral values and political attitudes. In the paper, a new computational approach from NLP, ML, and SNA is presented, whereby political opinions of issue-based discussions and their network structures are examined to determine the prevalence of like-minded spaces in terms of on their homogeneity versus heterogeneity. This procedure provides a more nuanced view, as it allows us to consider user communication in the dynamic opinion climate of different political issues. While P1 shows that RNNs achieve better performance with word embeddings, it also suggests that other ML methods using TF-IDF may yield better results, which is why P2 follows up on this suggestion and provides higher performance scores on these three datasets using SVM and TF-IDF. The results on the three controversial topics (adoption rights for homosexual couples, wearing religious headscarves, and climate change) illustrate a moderate level of opinion-based heterogeneity, which is characterized by the fact that users with positive or negative attitudes toward the topics tend to communicate in heterogeneous discussion spaces.

Furthermore, P2 not only considered homogeneity in the entire network with all videos and comments (macro level), but also identified community structures with the help of the fast-greedy algorithm to divide the network into smaller sub-networks and thus consider homogeneity from the micro level. The segmentation of the subnetworks enables a detailed view of network structures, in particular the communication of individual communities that are connected to influential hubs (channel owners). Likewise, heterogeneous communication behavior within the sub-networks examined was measured using the E-I index. This second result demonstrates the considerable relevance of specifically elaborating the community structures in networks to draw comparisons between the macro and micro level. This distinction between macro level and micro level may help us to understand complex communication networks such as YouTube from different perspectives in order to gain a deeper understanding of how opinions in the individual communities and their hubs are connected and disseminated, given that it is only through their interconnectedness with other communities that a global network emerges that addresses and covers a unified topic. These different aspects are particularly important for the discovery of homogeneous spaces, since the macro level can show an overall picture of how users discuss across different communities, while the micro level allows a detailed perspective of the individual communities and their opinions in order to better understand their users and internal structures. Thus, bridging the macro and micro levels from the network perspective can reveal different relationships between communication patterns, an aspect often neglected in science.

Although the results of Paper P2 revealed a heterogeneous prevalence of political issues, previous studies have suggested that homogeneous communication clusters may occur among marginal groups that share extreme values. Paper P3 argues that for the study of opinion-based homogeneity of conspiracy theories, it is reasonable to employ the spiral of silence theory, since it can be assumed that people who support the conspiracy theory see themselves as a minority in society and for this reason are more likely to communicate with like-minded people. The majority opinion would thus be viewed by opponents of the conspiracy theory who refute the theory and present facts. Accordingly, Paper P3 examined videos and comments relating to three conspiracy theories (Hollow Earth, Chemtrails, New World Order) on YouTube. In addition to the comments (N = 123,642), one objective was to analyze the content of the video (N = 176) in order to determine how dominant videos that likely support or debunk the conspiracy theory are on the platform. Reactions (likes, dislikes) to the respective videos also

served a crucial purpose by providing social feedback regarding interactions. To determine the prevalence of opinion-based homogeneity in the network, the videos and their user comments were examined. As the technology around artificial intelligence is in transformation and research on the creation of ML and DL models is constantly evolving, the BERT model allows us to draw on a powerful methodological approach that is even more suitable for text classification and is also suitable for training small datasets. Regarding the spiral of silence theory and the fragmentation of marginalized groups, the analysis generally shows how the opinion climate on conspiracy theory content evolves as people with different opinions interact online. The most important result of the analysis in relation to the spiral of silence theory is that homogeneous communication patterns exist between YouTube users who support the existence of conspiracy theories. Table 5 lists the most important parameters and results once again, summarizing the findings of Papers P2 and P3 on opinion-based homogeneity. The table also includes the context, topics, evaluated classes, and calculations of the class and global E-I index. The results in Table 5 on Paper P2 indicate that the discussions on political and controversial topics, i.e., adoption rights, headscarf ban, and climate change, are characterized more heterogeneously as the global and class E-I index are in a positive range of values. If one considers the results in Paper P3, in which the three different conspiracy theories, Hollow Earth, Chemtrails, and New World Order, were examined, one sees that in particular users who support the conspiracy theories communicate in a more homogeneous discussion network. In contrast, in two out of three topical contexts, users who counter the conspiracy theories have a more heterogeneous communication network, with the exception of the conspiracy theory of Chemtrails, whereby the class E-I index is zero and is thus neither homogeneous nor heterogeneous. More precisely, this means that this is a balanced communication relationship in which an equal number of proponents and opponents of the conspiracy theory communicate.

Table 5. Comparison of the prevelance of opinion-based homogeneity

Homogeneity	Context	Topics	Classes	Class E-I	Global E-I
Opinion-based (P2, P3)	Politics	Adoption rights	Positive	0.74	0.72
			Negative	0.70	
		Headscarf ban	Positive	0.78	0.58
			Negative	0.52	
		Climate change	Positive	0.76	0.61
			Negative	0.52	
		Hollow Earth	Pro-theory	–0.785	0.118

	Conspiracy theories		Contra-theory	0.708	
		Chemtrails	Pro-theory	–0.221	–0.131
			Contra-theory	0.031	
		New World Order	Pro-theory	–0.549	–0.549
			Contra-theory	0.377	

Another result of P3 also points out the strong distribution of conspiracy theory videos advocating the existence of the theory compared to debunking videos. Focusing on the labeled user-generated comments and replies, the findings showed more comments supporting the Chemtrails and New World Order theories than those debunking them. The conspiracy theory about Hollow Earth, however, did not yield this outcome, since there are more counter comments than comments supporting the theory. In this context, Paper P3 raises concerns that YouTube's recommendation algorithm may encourage the spread of misinformation on the platform, as it suggests additional content, thereby shaping and reinforcing users' opinions.

In summary, from the results of P2 and P3 on the investigation of opinion homogeneity, it can be said that the different results could be explained based on the context and the user group. While social issues in P2 ensure that people with different perspectives on a topic discuss it more heterogeneously, it is noticeable that this is not the case in P3 with topics such as conspiracy theories, which represent a marginalized group for advocates, and that these people tend to fall back into more homogeneous communication channels and interact with like-minded SNS users.

While P2 and P3 showed mixed results on the extent of opinion homogeneity, it is unclear when answering the question to what extent this also applies to informational homogeneity, since the focus here is not on opinions but on information and facts. Papers P4 and P5, therefore, address this aspect. To this end, Paper P4 introduces the concept of informational homogeneity. To identify the prevalence of informational homogeneity and misinformation in communication networks on YouTube and how they are interconnected, Paper P4 examined this analysis by studying the current COVID-19 pandemic over three months (January–March 2020). As previous research often overlooks the temporal comparison to the evolution of the opinion climate, P4 argues that especially at the beginning of the COVID-19 pandemic, when a lot of misinformation was being spread on the network, a temporal analysis of this informational homogeneity is a relevant point to better assess the progression of information and the views it

expresses. In the analysis of homogeneity, a total of 2,585,367 comments and 10,724 videos were analyzed, and a sample of these data was annotated according to the classes of misinformation and non-misinformation. In general, the same methodological procedure was used as in P3, which corresponds to a combination of NLP, BERT, and SNA. One difference to the previous Papers P2 and P3 is that in P4, a distinction is made between two networks, one involving the network of videos, comments, and replies, and the other consisting of comments and replies. Nevertheless, both types of network show similar outcomes. Furthermore, the findings suggest that there is a small amount of misinformation with different variations on YouTube, but these are exchanged in a very dense heterogeneous information network and, thus, one cannot assume specific fragmented subgroups. Regarding the temporal effects over the three months, only minor deviations to the prevalence of informational homogeneity were identified, whereby users spreading misinformation were in a constantly highly heterogeneous information environment.

Whereas the previous Papers P2–P4 focused on the communication networks of comments to analyze the active communication behavior of user opinions and information, P5 is about the study of YouTube's recommendation system with regard to politically related videos in order to analyze the functioning of algorithms and their proposed content. The controversy surrounding the existence of filter bubbles on YouTube is still a highly debated topic in academia and has led to differing results in previous research. In this regard, P5 argues that an essential component of understanding filter bubbles may be to analyze the content and related videos for their homogeneity/heterogeneity. Given the results of Paper P3, which reveals that more homogeneous communication behavior can exist among marginalized groups, P5 examines the homogeneity of the YouTube recommendation system, focusing the analysis on multiple levels (first depth, second depth) of recommended videos to politically neutral videos and right-wing populist videos in the network. Overall, the results of the study indicate that YouTube's recommendation algorithm follows a homogeneous pattern that ensures that similar content from right-wing populist videos, as well as neutral videos, is connected to the user. More precisely, the findings demonstrated that when starting a right-wing populist video, there is a 54% probability that the next video (in depth 1) will contain more right-wing populist information. However, the probability of encountering a right-wing populist video decreases to 37.7% in one of the next recommender depths. There is also a high degree of homogeneity among the neutral videos, which, however, suggest further neutral videos; only about 2% of the videos link to right-wing populist content.

Table 6 summarizes the results of Papers P4 and P5 on informational homogeneity and contains the same tabular characteristics as Table 5. The results of P4 illustrate that users who disseminate misinformation find themselves in a heterogeneous information network, as highlighted by the high positive E-I index in the months of January, February, and March. In contrast, the non-informational class exhibits strongly homogeneous behavior, as a negative E-I index was calculated in all months. These high negative values can be explained by the binary annotation, since any information about COVID-19 that does not consist of misinformation is represented here, and thus the proportion is also substantially higher. The previously mentioned results of P5 on homogeneity also become more apparent when examining the values of the E-I Index in the table. The global E-I index has a negative value for both the initial right-wing network and the initial neutral network, meaning that the recommendations of the system show homogeneous behavior with regard to political videos.

Considering the results of both papers, it is striking that similarities exist in relation to the results of P2 and P3, in which opinion-based homogeneity was investigated. More specifically, the findings reveal a thematic consistency, as communication about general topics such as health or climate change is characterized by heterogeneity, while topics associated with extreme political attitudes or conspiracy theories tend to be characterized by homogeneity.

Table 6. Comparison of the prevelance of informational homogeneity

Homogeneity	Context	Issues	Classes	Class E-I	Global E-I
Informational-based (P4, P5)	COVID-19	January	Misinformation	0.788	–0.508
			Non-misinformation	-0.795	
		February	Misinformation	0.842	-0.587
			Non-misinformation	-0.850	
		March	Misinformation	0.839	-0.652
			Non-misinformation	-0.869	
	Politics	Initial Right-wing	Neutral	/	-0.337
			Right-wing	/	
			Other	/	

	Initial Neutral	Neutral	/	-0.508
		Right-wing	/	
		Other	/	

4.2. Influencing factors and their opinion dynamics

In line with these aspects, paper P6 investigates the influence of opinion leaders based on the psychological concept of attitudinal ambivalence in order to analyze the opinion climate using ABM. More specifically, the authors are interested not only in the general influence of opinion leaders, but also in how the opinion climate behaves when they hold an ambivalent opinion, as well as when they discredit the opposing opinion group and thus strengthen their own position. For a more realistic picture of the impact of influential players, P6 analyzed two different opinion camps, in which different distributions of opinion leaders in the respective opinion camps (in this case red and blue) operate in the network and disseminate their opinions. The investigation of opinion leaders is therefore an important factor for observing the emergence of and changes in homogeneity with regard to the opinion climate and for drawing conclusions from this. Here, P6 develops a dynamic opinion model that builds on the DeGroot Model and considers two network topologies to account for the social influence of the connected neighbors in opinion formation. The model differs most notably in the update function, which ensures that agents update their opinions based on their neighbors at each tick. In this ABM, opinions are represented two-dimensionally, meaning that each agent has two values for the respective opinion group (red = 0.5, blue = 0.8); if an agent has the same values, they are ambivalent.

One of the key findings of the paper shows that even a small number of opinion leaders in a network environment have an influence on the opinion climate. The findings indicate, moreover, that an unequal distribution of opinion leaders from the two camps leads to an unbalanced opinion climate in which the opinion camp is dominated by the higher number of opinion leaders from the respective camp. However, if both opinion groups are represented by the same number (5 vs. 5) of opinion leaders, a balanced opinion climate is achieved. Another result shown by paper P6 is that ambivalent opinion leaders can lead to an increase in the number of ambivalent opinions in the network. Thus, it shows that the more moderately the opinion leaders spread both opinions in the network, the higher the degree of network ambivalence is. Furthermore, the results show that a network with only ambivalent opinion

leaders (i.e., red: 0, blue: 0, ambivalent: 12) leads to a balanced opinion climate, whereby the agents are influenced equally with ambivalent opinions. However, further insights also revealed that users were less ambivalent if, in addition to the ambivalent opinion leaders, additional opinion leaders (red: 0, blue: 12, ambivalent: 25) from other opinion groups were represented, as these had a stronger influence on the formation of opinions in the network.

In the discrediting scenario, in addition to the stepwise adjustment of the discrediting value of one particular opinion camp, the distribution of opinion leaders (i.e., 1 vs. 1, 5 vs. 5, 12 vs. 12, 25 vs. 25 and 50 vs. 50) was also taken into account. The distribution of the different opinion leaders of the respective opinion camps was equal in order to provide a comparable and balanced comparison. The findings in this regard have shown that there is a strong effect on winning over the opinion climate if opinion leaders ensure that the other opinion is discredited. To win the opinion climate for an opinion leader, only a few arguments against the opposing position are sufficient, even if additional opinion leaders of the camp are present. Furthermore, the findings demonstrated that the more arguments against the other opinion group are disseminated, the smaller the number of agents in favor of the discredited opinion. Similar to the previously mentioned results, it was found that with a critical mass of at least 12 opinion leaders in each opinion camp, opinion leaders who spread discrediting opinions have the strongest impact on winning the opinion climate. In summary, the opinion leader concept with its different forms (univalent, ambivalent, discrediting) can be seen as an influencing factor that could be responsible for the emergence as well as the change of homogeneity under certain conditions and, thus, can change the opinion climate over time. As an outlook for further research, paper P6 illustrates that not only would it be helpful to adapt this model to real network structures of social media, but suggests that this model can also be used for further research on the spiral of silence to analyze individual opinion expression.

In this context, Paper P7 demonstrates the impact of the community structure of networks to identify how they influence opinion formation. The spiral of silence served as a theoretical derivation to implement its principles in the applied agent-based model and builds on the work of Ross et al. (2019). In the newly applied model, no preferential attachment model is applied, but a stochastic block model, which makes it possible to investigate the spiral of silence effects in the communities, the size and number of communities, as well as the interconnectedness of the communities.

A core insight of Paper P7 is that the fragmentation of very many small communities facilitates the ability of agents to have minority opinion represented by a larger portion of the overall population. This encourages agents within a densely networked community to voice their opinion, even if it is a minority opinion, because it is insulated from other communities and the silencing process is focused on the respective individual communities locally. As the authors carefully formulate and leave open to interpretation, these virtual homogeneous spaces of minority opinions, where the global consensus does not correspond to the majority opinion, can also have negative consequences, such as radicalization in subgroups. Another finding that the authors of paper P7 point out is that a higher level of connectivity between communities ensures that the mechanism of the spiral of silence can emerge across all communities, which diffuses throughout the entire network. The process of the spiral of silence, through this increased connectedness between communities, can create pressure not to express opinions—even when there is a silent majority of other agents who share them. Figure 5 shows how the number of communities has an impact on a minority's ability to keep expressing their opinion.

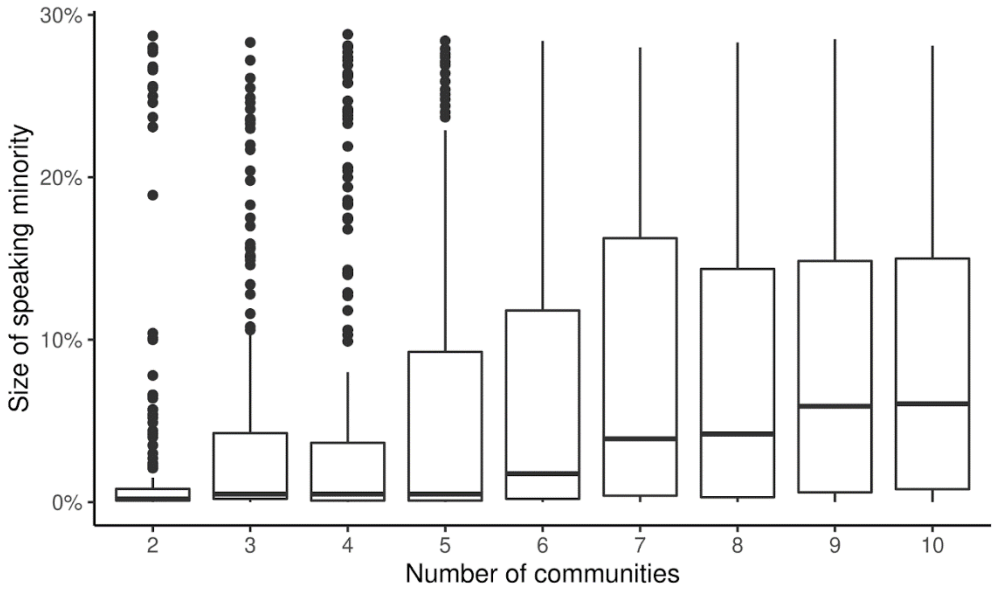

Figure 4. The impact of community numbers on a minority's ability to continue expressing its views (P7)

5. Discussion and implications

The joint consideration and an overarching discussion of the various results of the individual research papers helps to achieve a more global view of the two research questions investigated in the dissertation. In the following sections, the two aspects of the prevalence of opinion-based and informational homogeneity (Sect. 5.1), as well as the influencing factors that change the homogeneity in networks (Sect. 5.2) are discussed in depth. Sections 5.3 and 5.4 present implications for further research as well as for practice. The chapter ends with Section 5.5, which presents the limitations of the dissertation and discusses further aspects for future research.

5.1. Prevalence of opinion-based and informational homogeneity

The first research question asked how high the prevalence of opinion-based and informational homogeneity is. To answer this question, the cumulative dissertation presented five papers (P1–P5). P1 focused on the methodological comparison of RNNs and traditional ML models that have been trained in combination with word embeddings. Thus, Paper P1 does not directly address the investigation regarding homogeneity in social networks—it serves much more as an important foundation on which to build a deeper understanding and acceptance of social media data, such that it can be subsequently analyzed using NLP and ML methods, and this acquired knowledge can be taken up in the papers. In particular, processing data from social media still poses a great challenge, as it is not only noisy (Stieglitz et al., 2018), but its context also needs to be understood in order to further process it efficiently with algorithms (Pustejovsky & Stubbs, 2013). Furthermore, the paper also shows the relevance of comparing the performance of different ML models in order to better evaluate them afterwards, as was carried out in papers P2–P4.

When looking at the results of P2 and P3 on opinion-based homogeneity on YouTube, clear differences emerge that are very much related to the context and thus the topic. The results of paper P2, that is to say that more heterogeneous communication behavior was found among users with different attitudes toward the topic and that, thus, they do not manifest with like-minded people in political ideological groups, is contrary to previous findings investigating ideological homogeneity (Bakshy et al., 2015; Barberá et al., 2015; Del Valle & Bravo, 2018).. However, P2 is in line with a recently published study that looked at political tie building on Facebook (Cargnino & Neubaum, 2021). Having said that, it is important in this context to

distinguish that P2 focuses on active interaction with concrete discussions, while Cargnino & Neubaum (2021) focus on the process of forming digital connections in the form of friendships. They found that only a small proportion (36%) of users have ties to politically like-minded people and that the majority of users are exposed to different opinions. Hence, both studies address "connections" between people and their homogeneity, but on different levels. However, the findings of P2 are in contradiction with the conclusions of P3, which dealt with opinion-based homogeneity in the context of conspiracy theories. Here, the results suggest that there may indeed be homogeneous spaces of like-minded users who endorse a conspiracy theory. This result is consistent with previous findings that examined the active communication of conspiracy theories on Facebook and Reddit and demonstrated that they can spread in homogeneous communities (Del Vicario et al., 2016; Phadke et al., 2021). In particular, the findings concerning the platform Reddit for the investigation of conspiracy communities and their communication interactions revealed similar results; thus, the study found that future members who join a conspiracy community not only engage in similar conspiracy discussions, but also seek a direct exchange with other members of the conspiracy community (Phadke et al., 2021).

Examining the results of P4 and P5 on informational homogeneity reveals similar patterns to the results on opinion-based homogeneity. While the results of P4 focus on the prevalence of homogeneity in the discussion networks on the keyword "corona" over three months and show heterogeneous behavior of the discussion landscape, P5 assesses as homogeneous the recommendation behavior on the YouTube platform on neutral and right-wing populist videos, where the results show that there is indeed an increased likelihood that similar topics are further suggested. P5's results, along with comparable studies dealing with YouTube's recommendation system, likewise support the fact that users who have already consumed politically extreme or non-mainstream content such as conspiracy myths have a higher likelihood of accessing more extreme videos suggested by the algorithm (Faddoul et al., 2020; Hussein et al., 2020; O'Callaghan et al., 2013; Whittaker et al., 2021). The exploration of informational homogeneity additionally illustrates that communication can be more homogeneous for users of these fringe groups if they believe in the correctness of these attitudes. Accordingly, YouTube can be a place of homogenous communication, albeit for fringe groups.

Although there are no comparable studies to the results of P4 investigating the homogeneity on an active discussion network regarding the COVID-19 pandemic, previous studies have also

shown that there was already a large amount of misinformation on SNS during the pandemic (Bridgman et al., 2020; Cinelli et al., 2020; Kouzy et al., 2020).

The joint consideration of opinion-based and informational homogeneity reveals two interesting findings that address the answer to the first research question and thus make an important contribution to further research in political homogeneity. First, it shows that Bruns' assumptions (Bruns, 2019a, 2021) that political homogeneity (based on opinions or information) is not widespread among average consumers, as can be empirically confirmed by the applied methodological procedures in papers P2 and P4. The results of the studies indicate that users of YouTube develop more heterogeneous communication behavior when they exchange opinions on politically controversial topics or current information describing the current political situation (as in the case of COVID-19) and find a discourse with each other, whereby it is not relevant which position they hold on the topic. Indeed, previous research in this area has been able to show that political opinions in discussions or the use of news to obtain information tends to lead to a heterogeneous media landscape in political discussion networks (Brundidge, 2010), as people happen to be exposed to different views and opinions (e.g., (Kim, 2018; Lee et al., 2014; Lu & Lee, 2019; Vaccari et al., 2016)

On the other hand, the empirical results from P3 and P5 make an important contribution to investigating the fragmentation of extreme opinions and ideological ideas in addition to the political topics on YouTube, and thus to a better understanding of the fragmentation of certain groupings in the network and the avoidance of their radicalization. The results of P3 show that in particular the users of these groups (e.g., conspiracy theorists) have a high degree of homogeneous communication structures in the form of opinions and thus seek contact with like-minded people. Taking into account the results in paper P7, it has also been empirically shown that under certain circumstances, an alienation of the majority society from the global consensus of opinion takes place, since the users in the many small communities have little connectivity to other users in other communities. Thus, it can be concluded that not only the extreme topic can favor homogeneous spaces, but also the functionality of the platform in terms of proposing new content and especially how users can join in communities. Algorithm recommendations, in particular, can reinforce the formation of isolated communities (Santos et al., 2021) and create a sense of community belonging that can be evoked by even single interactions (Rotman et al., 2009). In this regard, according to researchers, homogeneity can become a critical driver of network interactions in conjunction with recommendation algorithms, as these ensure that community visibility is promoted or diminished, even if that group is a minority in the network

(Fabbri et al., 2020). In this way, the recommendation systems on platforms allow users to discover new political content that they were not formerly aware of through their previous sphere of interest (Munger & Phillips, 2022). This may involve the risk that users increasingly see extreme and fringe content on the network (Whittaker et al., 2021) and, on the other hand, that users already holding extreme attitudes have further extreme content suggested to them by the algorithms (Liu et al., 2021). In view of the previously highlighted challenges, a network study by Stern and Livan showed that the existence of many different opinions can make it difficult to form homogeneous spaces in which a common consensus prevails (Stern & Livan, 2021). These results highlight issues similar to those addressed in the dissertation, as within the investigated networks a large proportion of users who cannot be classified as belonging to marginalized groups have shown heterogeneous communication, and a large number of comments did not address the actual topic (topic-independent). The aforementioned efforts and problems related to the YouTube platform may be solved through further computational applications by not only addressing the proposed political videos as in P5, but by further investigating how the discussion network evolves to do so as in P4, in order to create a detailed image in which a bridge between algorithms and UGC in the form of comments and videos is built and also considered.

5.2. Influencing factors that change the homogeneity of opinions and information

While previous papers (P2–P5) implicitly took a snapshot of collected data from different topics to investigate homogeneity on the YouTube platform, the papers discussed here, P6 and P7, investigate the dynamic and temporal aspects of opinion leaders and community structures using ABM. The simulation of agent-based models offers the opportunity to simulate the temporal dynamics of network interactions in order to gain a deeper understanding of the processes of human behavior, such as the formation and dissemination of information and opinions in social networks. In this context, social science theories not only serve as an experimental object to discover relevant insights and factors, but also serve as theoretical foundations for a more realistic specification of parameters and rules in the modeling process. The observation of dynamic processes can also provide conclusions about existing network structures, as interactions on different topologies might be compared in order to gain insights into the connectedness of individuals. In this regard, the second research question asks what factors influence the emergence and change of homogeneity in a network of opinion and information as described, thus drawing conclusions about homogeneity. Although the

homogeneity of these papers was not directly determined by using the E-I index, conclusions can nevertheless be drawn from the dissemination of the opinion climate.

Paper P6 investigated the influence of opinion leaders in a two-dimensional opinion environment to determine how social network communication evolves in two different network topologies. For this purpose, three different types of opinion leaders were implemented: those who 1) adopt an attitude for their own camp, 2) adopt an ambivalent attitude and are modeled according to the theoretical basis of attitude ambivalence, and 3) those who adopt a discrediting attitude toward the other opinion camp and thus weaken it. Since these aspects have not been investigated in previous research, this paper uses a computational approach to highlight the significant role that opinion leaders play in disseminating opinions and influencing the opinion climate. As the results of P6 indicate, ambivalent opinion leaders can balance the opinion climate to a certain extent, as more users adopt an ambivalent opinion where there are differences in the two network structures evaluated. It is known from previous studies that different network topologies have effects on opinion dynamics in the network when interactions between opinions occur (Rodrigues & Da F. Costa, 2005). Since network topology is important for opinion dynamics and diffusion, it would be plausible that opinion leaders are placed at crucial points in networks to achieve a large reach so that many people meet them. With regard to the findings in Papers P2–P5, no explicit factors concerning the change in homogeneity were identified, as the studies did not consider dynamic processes and hence only provided conclusions about the prevalence of opinion-based and informational homogeneity in networks of topics at a certain point in time (snapshot). However, Paper P4, which takes into account the temporal component of the fragmentation of the information landscape by looking at it on a monthly basis, observed only a minimal difference in the overall homogeneity of information when influential hubs in the network are considered, or not, although these hubs were also not defined as opinion leaders, but as videos. Performing ABM, on the other hand, ensures the realization of various scenarios to be investigated, since the parameter combinations enable the change in influencing factors over time to be recognized. Thus, changes in homogeneity can be tracked by highlighting influencing factors that emerge during the course of the simulation, such as the number or distribution of opinion leaders. Furthermore, ABM is much easier to perform since, for example, ambivalent opinion leaders in the network can be defined based on theoretical derivations rather than determining them in real data sets by identification algorithms and annotating their opinions. However, future studies may extend to investigate the interplay of simulation with real-world data and develop an even more accurate ABM. This may involve the integration of real network data from online social platforms, where each

individual interaction is annotated for the characteristic features of users, (such as message sentiment or identification of opinion leaders), such that communication can be represented over time. This annotated dataset can thereby serve as a baseline and be used to conceptualize further samples for ABM. Much like a permutation test for networks, the edges can be rewired in order that different network structures can be included in the overall evaluation. It would also be conceivable that within this process, the E-I index is computed at intervals, meaning that in addition to the ratio of opinions, another indicator for the evaluation of homogeneity in the network can be considered, one which also considers the connections to each other.

In particular, since P7 considers the connectivity of nodes in the network, the findings indicate that many smaller communities build up a minority opinion over time when these users have few connections to other communities. In this respect, not only the identification of opinion leaders could be relevant, but rather the integration of ambivalent opinion leaders to balance the opinion climate as people within communities come into contact with ambivalent opinions on certain topics, thereby preventing exclusion. This process would lead to people encountering cross-cutting information and thus to heterogeneous opinions being exposed. This distribution could be especially important in the political and social context. However, these ambitions can also lead to the exact opposite, whereby a strong imbalance of opinion leaders to different opinion groups is established in the network. Particularly regarding the community structures in P7, these opinion leaders in the communities could win the opinion camp for themselves, as they are sealed off from other opinions by their low connectivity in the network, where the minority opinion does not even enter into a spiral of silence process. This is also in line with the results of previous research, which has shown that the minority is more likely to speak out if they consider themselves in a safe environment (Matthes et al., 2010).

It should also be noted that Paper P6 has reported that discrediting opinion leaders with respect to the other opinion group can further enhance this effect, making it easier to win over the opinion group. As P6 argues, this scenario would support the concerns of Sunstein, who argues that over time this will favor the polarization of the two opinion groups (Sunstein, 2017).
Since the modeling in P6 considers that agents are confronted with a diversity of opinions due to the two network structures investigated over time, the question remains open as to the extent to which a polarization effect could occur in strongly segmented networks. In favor of the polarization of two opinion groups, however, one would argue that the network splits into two strongly segregated networks in which people are only reinforced by opinions they perceive in

their environment. Opinion leaders might favor this process in the setting mentioned above. The results also show that a strongly unbalanced distribution of opinion leaders can lead to a situation in which the opinion climate is largely won over and the minority opinion is only weakly represented. This would suggest that opinion leaders can contribute to a higher degree of homogeneity in the network over time.

However, Paper P2, which is based on real-world data on political issues, shows that the vast majority of YouTube user comments have no real connection to the actual topic, and thus many opinions were annotated as off-topic. Although no opinion leaders could be identified in the results of P2, the general opinion climate was very balanced in terms of the number of positive and negative comments, which also led to heterogeneous communication behavior.

The results of P7 also show further behavior that leads to a global spiral of silence and which transcends several communities, as there is higher connectivity between the communities in this case. This would also suggest that a more heterogeneous opinion climate would develop until a consensus of the majority opinion was found, in which opinion leaders could possibly also have an influence on the opinion picture. This process of a global spiral of silence could reflect a closer picture of reality, given that people are confronted with a variety of information and also do not focus on just one network (Shearer & Gottfried, 2017).

Furthermore, the findings from P6 and P7 have also pointed out the need for the development of ABM in the field of CSS, where theoretical derivations from previous interdisciplinary research are used to develop and simulate new agent-based models to achieve a deeper understanding of the dynamics of opinion and its formation of social phenomena. Likewise, ABM findings can help to generate new hypotheses that can then be tested with field experiments.

5.3. Implications for research

The dissertation provides implications in the field of CSS and shows new ways to measure the homogeneity of communication channels compared to the previous ideological homogeneity. The methodological approach from NLP, ML, and SNA in P2–P5 can be used as a blueprint for research in the future to analyze political and controversial issues based on opinions and information according to their homogeneity in the network. The applied method and its implementation would also be suitable for carrying out cross-platform analysis, in which a cross-platform comparison of different topics is made possible; in this case, the network structures of the other platform would need to be evaluated and adapted, if necessary. Furthermore, it was shown that not only stationary, but also temporary changes in homogeneity can be measured (P4), in order to be able to trace a detailed and temporal development of different topics. Examining the temporal relationship between different events in more detail might help to gain a clearer understanding of the evolution of political topics. Especially in the current COVID-19 pandemic, these temporal comparisons can provide important insights, as they can specifically trace whether certain events or political precautions have led to a change in opinion on topics (e.g., vaccination, mandatory masking). In addition, the results showed that other aspects such as opinion leaders and community structure have a possible influence on the expression of opinions in the network (P6, P7). Research on misinformation may also adopt this approach to identify informationally homogeneous clusters and opinion leaders in order to detect the spread of false information as soon as possible, thereby enabling the initiation of countermeasures and corrections to address misinformation. As studies have already shown, countermeasures such as warnings and additional explanations would enable users to counter misinformation, as they can form their own opinions (Kirchner & Reuter, 2020).

The dissertation also provides further theoretical added value to the study of marginal groups (extreme ideologies, conspiracy theories) to better understand their communication networks and shared content (P3, P5). Previous studies have already suggested that these marginal groups may be more likely to have users in more homogeneous spaces and only draw their information from these users. The results of Paper P3 suggest that users who support a conspiracy theory (and thus can be considered a marginal group) have more homogeneous interactions of discussions on content and comments that are consistent with their opinions, and thus are consistent with the silence of spiral theory. Likewise, platforms such as YouTube, which operate with a recommendation system based on personalized user data, could further amplify this effect, whereby users from marginal groups increasingly view like-minded content that

arguably aligns with their worldview (P5). The YouTube platform, on which there is currently still a low level of research, demonstrates through its diversity of functions and the abundant UGC a very broad spectrum of analysis possibilities to better understand social phenomena and the way content is suggested by algorithms. Furthermore, the use of ABM, in which more complex social psychological processes are represented, can produce meaningful findings that can inform current social debates. Overall, the dissertation not only provides a methodological blueprint, but also a theory-driven blueprint, focusing on how social science theories can be used with computational methods to investigate current phenomena, thus promoting the interdisciplinary integration of knowledge.

5.4. Practical implications

The results of this dissertation show that the practical implications for political communication can be addressed from different perspectives. From the perspective of political marketing, i.e., the strategic positioning and communication of political entities with their environment (Lock & Harris, 1996), the computational approaches in the papers (P2–P4) have practical relevance for political parties. Since political parties present their election programs and campaigns on social media and thus use them as a marketing tool (Cameron et al., 2016; Enli, 2017), these methods could be used to determine and subsequently analyze the opinion-based homogeneity on specific political topics of user-generated comments. For instance, political parties may use data from social networks for their campaigns to gain a deeper understanding of how homogeneous/heterogeneous users' opinions and information on certain topics are (e.g., national leadership candidates). However, not only is the content relevant, but also how users are connected in the network. Bringing relevant messages to voters quickly and efficiently also requires identifying opinion leaders for campaigns who can influence target groups through their strong ties (Ozturk & Coban, 2019). The findings from P6 and P7 have demonstrated how opinions and information are distributed in networks depending on influencing factors such as opinion leaders and their community. Through simulations based on real network data, scenarios can be modeled that help to identify opinion leaders for campaigns and work out strategies to spread opinions in a targeted way in the network, whereby the network structure can also be taken into account in order that various platforms are represented.

The dissertation highlights another practical implication in relation to political education. The COVID-19 pandemic has demonstrated how rapidly misinformation can spread on the network and how vulnerable individuals are if they unquestioningly trust misinformation (Cinelli et al.,

2020; Lazer et al., 2018; Melki et al., 2021). In addition, polls revealed that the social media usage of German Internet users has grown due to COVID-19, with 62% consuming more content and 28% sharing more posts about current events (Bitkom e.V., 2020). The study of informational homogeneity might be a useful method to identify homogeneous clusters of misinformation and influential hubs in the network in order for platform providers to initiate countermeasures. However, as these measures only solve a short-term problem, long-term measures should be sought within political education in public administration to sensitize the young generation to the current media landscape and media use, so that they can learn to deal critically with the platforms and better assess information (Milbradt & Hohnstein, 2017). Education on how this misinformation is disseminated in online networks and what content it contains could deepen the opinion-forming process of young people and reduce dangers in social media.

5.5. Limitations and future directions

As a first limitation, the investigation of homogeneity on only the YouTube platform needs to be emphasized. This dissertation examined most studies regarding the YouTube social media platform on different topics and research areas, thereby highlighting the significance of the various studies on the prevalence of homogeneity in opinions and information. Indeed, the platform-specific study of one platform shows its strengths in the detailed comparability of the prevalence of topic homogeneity; however, the question remains as to what extent the prevalence on these topics is also represented on other platforms (e.g., Facebook, Twitter, Instagram, or TikTok). Assuming that in a cross-platform comparison, the same thematic content is studied on different platforms, an important aspect from a scientific point of view would be to highlight the differences in terms of homogeneity or opinion dynamics in order to obtain a more global overview. It is know from previous research that people do not prefer just one source of information, but are active on, and also combine, different platforms (Newman et al., 2021; A. Smith et al., 2018). Previous evidence also suggests that political communication differs by social media platform (Stier et al., 2018; Valenzuela et al., 2018; Yarchi et al., 2021). This differentiation in political communication might be connected to technical affordances in social media, i.e., how users interact with the platform in terms of its implemented functionality and architecture (Bossetta, 2018). For example, Facebook only displays feeds from friends or groups that one personally follows, whereas Twitter exposes one to external information such as retweets from followers. Another potential issue that may cause communication to differ across platforms is referred to as persistence (accessibility of information), which, along with

replicability, scalability, and searchability, is another affordance in network communication (boyd, 2010). Studies on this have found evidence that message persistence has an impact on people's behavior in that it reduces willingness to express political attitudes associated with higher individual costs and lower personal benefits (Neubaum, 2021). To investigate these cross-platform comparisons, it might be useful to consider the issues raised here when evaluating the results in relation to political communication. However, studying different platforms can yield a deeper and more global understanding of political communication (Garrett et al., 2012), while at the same time providing further perspectives on virtual homogeneous spaces by comparing the use of individual users in terms of their media behavior (Dubois & Blank, 2018). This would also counteract the "platform bias" currently prevalent in the literature, whereby the vast majority of research focuses exclusively on the Twitter platform (Van Aelst et al., 2017).

However, this may be related to the fact that data access to the platform-specific API is more readily available than for other platforms, which may also be linked to further restrictions and hurdles to data access (Stieglitz et al., 2014). Furthermore, the aspect of "Big Data" should not be overlooked, as the analysis of a multitude of social media data on different platforms poses a far greater challenge in the areas of data discovery, data collection, and data preparation (Stieglitz et al., 2018). Particularly since SNS have become an increasingly important part of human life and have changed the way people communicate with each other, it is even more important for CSS to analyze diverse data types (textual, visual) in future research to gain new insights in this area. For example, visual framing techniques could be used to answer questions about how visual information such as YouTube thumbnails are displayed and how they influence users. Here, future research might use computer vision, which is the automatic analysis of visual information such as images or videos in combination with artificial intelligence to derive meaningful information from the data in order to identify patterns. This raises not only ethical issues, but also questions about data protection and how science should deal with publicly available personal data on SNS.

From the point of view of CSS, which studies the social behavior of people by means of computational methods, it is for these reasons that new innovative concepts that overcome the previous barriers to reproducibility, transparency, and the basic ethical concepts of these data need to be developed, in order to provide clarity and to comply with the basic guidelines of modern science (Merton, 1973). The open science movement clearly demonstrates the

importance of researchers having a common repertoire to share data. This challenge, however, also goes hand in hand with the regulations of the platform providers, which, for the most part, prohibit researchers from sharing the data, thereby posing a major problem from the perspective of science, given that it is therefore not possible to establish the external replicability of this proprietary data (Theocharis & Jungherr, 2021). Another limitation that must be considered in the dissertation is that although it is technically possible to provide software to retrieve data from the platform, this might lead to slightly different results, since data may have been deleted by the platform or the user.

Studies that address ABM and use theory-driven research to examine social phenomena provide an opportunity to generate new data and hypotheses for future research. Thus, using P6 and P7, it was possible to illustrate the influence opinion leaders might have on opinions and the influence communities would have in expressing opinions, given the spiral of silence mechanism. Nevertheless, there is a principal limitation in simulation studies, as they generate an artificially created scenario that is supposed to represent a certain image of reality. The findings of the studies generally provide a basis for creating further hypotheses, which can be investigated through further research using field experiments and surveys to build a bridge between micro and macro levels. Thus, models that include opinion dynamics can provide insight into certain theories under pre-programmed conditions, making it possible to understand the micro and macro level relationship in more depth. In addition to examining individual sociological theories, future research may address how the interplay of multiple intertwining theories has an impact on opinion formation and homogeneity, as these theories might provide more accurate representations of reality, but are also more complex to model.

Conclusions on the two guiding research questions can be presented as follows: The results show that political homogeneity is not widespread for the average user and that YouTube users develop more heterogeneous communication behaviors when sharing (as in the case of COVID-19) about politically controversial topics or current information about the current political situation. However, the results also show that marginalized groups in society, such as people who advocate conspiracy theories, exhibit moderate levels of homogeneity and are more likely to engage with like-minded people. Furthermore, the results show that opinion leaders as well as the structure of communities in networks can be characterized as influencing factors in the change and emergence of political homogeneity, as they have an impact on opinion over time.

In retrospect, it can be summarized that new technological capabilities of computational methods can enable upcoming research to look at social phenomena from different perspectives, thus reducing the acceptance of society in relation to social media. This is also favored by the rapidly growing social media ecosystem, which promotes interactions between (new) platforms and access to more data. The upcoming challenge can be overcome by the interdisciplinarity of research fields to obtain a synthesis of knowledge from different disciplines. In this regard, the dissertation has presented results to expand our understanding of the prevalence of homogeneity of social networks, an understanding that needs to be deepened in further research.

References

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, *6*(2), 205316801984855. https://doi.org/10.1177/2053168019848554

Allington, D., Duffy, B., Wessely, S., Dhavan, N., & Rubin, J. (2020). Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. *Psychological Medicine*, 1–7. https://doi.org/10.1017/S003329172000224X

An, J., Kwak, H., Posegga, O., & Jungherr, A. (2019). Political Discussions in Homogeneous and Cross-Cutting Communication Spaces. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*(01), 68–79.

Armitage, C. J. (2003). Beyond attitudinal ambivalence: Effects of belief homogeneity on attitude-intention-behaviour relations. *European Journal of Social Psychology*, *33*(4), 551–563. https://doi.org/10.1002/ejsp.164

Auxier, B., & Anderson, M. (2021). Social media use in 2021. *Pew Research Center*.

Bachmann, R., Kemper, G., & Gerzer, T. (2014). *Big Data - Fluch oder Segen? Unternehmen im Spiegel gesellschaftlichen Wandels* (1. Aufl). mitp.

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 258–265. https://doi.org/10.1109/ASONAM.2018.8508646

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130. https://doi.org/10.1126/science.aaa1160

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. *Proceedings of the 21st International Conference on World Wide Web - WWW '12*, 519–528. https://doi.org/10.1145/2187836.2187907

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, *10*(2), e0118093.

Bessi, A., Zollo, F., Del Vicario, M., Puliga, M., Scala, A., Caldarelli, G., Uzzi, B., & Quattrociocchi, W. (2016). Users polarization on Facebook and Youtube. *PloS one*, *11*(8), e0159641.

Bitkom e.V. (2020). *Social-Media-Nutzung steigt durch Corona stark an* [Available Online]. https://www.bitkom.org/Presse/Presseinformation/Social-Media-Nutzung-steigt-durch-Corona-stark-an

Bliuc, A.-M., Smith, L. G. E., & Moynihan, T. (2020). "You wouldn't celebrate September 11": Testing online polarisation between opposing ideological camps on YouTube. *Group Processes & Intergroup Relations*, *23*(6), 827–844. https://doi.org/10.1177/1368430220942567

Bode, L. (2016). Political News in the News Feed: Learning Politics from Social Media. *Mass Communication and Society*, *19*(1), 24–48. https://doi.org/10.1080/15205436.2015.1045149

Bode, L., & Vraga, E. K. (2015). In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. *Journal of Communication*, *65*(4), 619–638. https://doi.org/10.1111/jcom.12166

Bond, R., & Messing, S. (2015). Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook. *American Political Science Review*, *109*(1), 62–78. https://doi.org/10.1017/S0003055414000525

Borge Bravo, R., & Esteve Del Valle, M. (2017). Opinion leadership in parliamentary Twitter networks: A matter of layers of interaction? *Journal of Information Technology & Politics*, *14*(3), 263–276. https://doi.org/10.1080/19331681.2017.1337602

Bossetta, M. (2018). The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election. *Journalism & Mass Communication Quarterly*, *95*(2), 471–496. https://doi.org/10.1177/1077699018763307

Boulianne, S. (2020). Twenty Years of Digital Media Effects on Civic and Political Participation. *Communication Research*, *47*(7), 947–966. https://doi.org/10.1177/0093650218808186

Boutyline, A., & Willer, R. (2017). The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology*, *38*(3), 551–569.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications*, *10*(1), 7.

boyd, D. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In *A networked self* (S. 47–66). Routledge.

Bozdag, E., & van den Hoven, J. (2015). Breaking the filter bubble: Democracy and design. *Ethics and Information Technology*, *17*(4), 249–265. https://doi.org/10.1007/s10676-015-9380-y

Brennen, J. S., Simon, F., Howard, P. N., & Nielsen, R. K. (2020). Types, sources, and claims of Covid-19 misinformation. *Reuters Institute*, *7*.

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., & Zhilin, O. (2020). The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-028

Bright, J. (2016). Explaining the Emergence of Echo Chambers on Social Media: The Role of Ideology and Extremism. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2839728

Bright, J. (2018). Explaining the Emergence of Political Fragmentation on Social Media: The Role of Ideology and Extremism. *Journal of Computer-Mediated Communication*, *23*(1), 17–33. https://doi.org/10.1093/jcmc/zmx002

Bruch, E., & Atwell, J. (2015). Agent-Based Models in Empirical Social Research. *Sociological Methods & Research*, *44*(2), 186–221. https://doi.org/10.1177/0049124113506405

Bruguera, C., Guitert, M., & Romeu, T. (2019). Social media and professional development: A systematic review. *Research in Learning Technology*, *27*(0). https://doi.org/10.25304/rlt.v27.2286

Brundidge, J. (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), 680–700. https://doi.org/10.1111/j.1460-2466.2010.01509.x

Bruns, A. (2017). Echo chamber? What echo chamber? Reviewing the evidence. *6th Biennial Future of Journalism Conference (FOJ17)*.

Bruns, A. (2019a). *Are filter bubbles real?* Polity Press.

Bruns, A. (2019b). *It's not the technology, stupid: How the `Echo Chamber' and `Filter Bubble' metaphors have failed us*. https://eprints.qut.edu.au/131675/

Bruns, A. (2021). Echo chambers? Filter bubbles? The misleading metaphors that obscure the real problem. In M. Pérez-Escolar & J. M. Noguera-Vivo, *Hate Speech and*

Polarization in Participatory Society (1. Aufl., S. 33–48). Routledge. https://doi.org/10.4324/9781003109891-4

Bruns, A., Burgess, J., Highfield, T., Kirchhoff, L., & Nicolai, T. (2011). Mapping the Australian Networked Public Sphere. *Social Science Computer Review*, *29*(3), 277–287. https://doi.org/10.1177/0894439310382507

Bryant, L. V. (2020). The YouTube Algorithm and the Alt-Right Filter Bubble. *Open Information Science*, *4*(1), 85–90. https://doi.org/10.1515/opis-2020-0007

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, *20*(1), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

Burgess, J., & Green, J. (2018). *Youtube: Online video and participatory culture* (Second edition). Polity Press.

Cameron, M. P., Barrett, P., & Stewardson, B. (2016). Can Social Media Predict Election Results? Evidence From New Zealand. *Journal of Political Marketing*, *15*(4), 416–432. https://doi.org/10.1080/15377857.2014.959690

Cargnino, M. (2020). The Interplay of Online Network Homogeneity, Populist Attitudes, and Conspiratorial Beliefs: Empirical Evidence From a Survey on German Facebook Users. *International Journal of Public Opinion Research*, edaa036. https://doi.org/10.1093/ijpor/edaa036

Cargnino, M., & Neubaum, G. (2021). Are We Deliberately Captivated in Homogeneous Cocoons? An Investigation on Political Tie Building on Facebook. *Mass Communication and Society*, *24*(2), 187–209. https://doi.org/10.1080/15205436.2020.1805632

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, *19*(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), 16598. https://doi.org/10.1038/s41598-020-73510-5

Cioffi-Revilla, C. (2010). Computational social science: Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 259–271. https://doi.org/10.1002/wics.95

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication*, *64*(2), 317–332.

Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., & Flammini, A. (2011). Political Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1). https://ojs.aaai.org/index.php/ICWSM/article/view/14126

Conroy, M., Feezell, J. T., & Guerrero, M. (2012). Facebook and political engagement: A study of online political group membership and offline political engagement. *Computers in Human Behavior*, *28*(5), 1535–1546. https://doi.org/10.1016/j.chb.2012.03.012

Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J., & Stefanidis, A. (2015). Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems*, *53*, 47–64. https://doi.org/10.1016/j.compenvurbsys.2014.11.002

Del Valle, M. E., & Bravo, R. B. (2018). Echo chambers in parliamentary Twitter networks: The Catalan case. *International journal of communication*, *12*, 21.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

Dijk, T. A. V. (1995). Discourse, Opinions and Ideologies. *Current Issues In Language and Society*, *2*(2), 115–145. https://doi.org/10.1080/13520529509615438

Donzelli, G., Palomba, G., Federigi, I., Aquino, F., Cioni, L., Verani, M., Carducci, A., & Lopalco, P. (2018). Misinformation on vaccination: A quantitative analysis of YouTube videos. *Human Vaccines & Immunotherapeutics*, *14*(7), 1654–1659. https://doi.org/10.1080/21645515.2018.1454572

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society*, *21*(5), 729–745.

Dubois, E., & Gaffney, D. (2014). The Multiple Facets of Influence: Identifying Political Influentials and Opinion Leaders on Twitter. *American Behavioral Scientist*, *58*(10), 1260–1277. https://doi.org/10.1177/0002764214527088

Dvir-Gvirsman, S. (2017). Media audience homophily: Partisan websites, audience identity and polarization processes. *New Media & Society*, *19*(7), 1072–1091. https://doi.org/10.1177/1461444815625945

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational Social Science and Sociology. *Annual Review of Sociology*, *46*(1), 61–81. https://doi.org/10.1146/annurev-soc-121919-054621

Ellison, N. B., & Boyd, D. M. (2013). *Sociality Through Social Network Sites* (W. H. Dutton, Hrsg.; Bd. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199589074.013.0008

Enli, G. (2017). Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election. *European Journal of Communication*, *32*(1), 50–61. https://doi.org/10.1177/0267323116682802

Evans, M. (2016). Information dissemination in new media: YouTube and the Israeli–Palestinian conflict. *Media, War & Conflict*, *9*(3), 325–343. https://doi.org/10.1177/1750635216643113

Fabbri, F., Bonchi, F., Boratto, L., & Castillo, C. (2020). The effect of homophily on disparate visibility of minorities in people recommender systems. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 165–175.

Faddoul, M., Chaslot, G., & Farid, H. (2020). A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos. *arXiv:2003.03318 [cs]*. http://arxiv.org/abs/2003.03318

Feezell, J. T., Wagner, J. K., & Conroy, M. (2021). Exploring the effects of algorithm-driven news sources on political behavior and polarization. *Computers in Human Behavior*, *116*, 106626. https://doi.org/10.1016/j.chb.2020.106626

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2018). Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 913–922. https://doi.org/10.1145/3178876.3186139

Garrett, R. K., Bimber, B., De Zúñiga, H. G., Heinderyckx, F., Kelly, J., & Smith, M. (2012). Transnational Connections| New ICTs and the Study of Political Communication. *International Journal of Communication*, *6*, 18.

Garrett, R. K., & Weeks, B. E. (2017). Epistemic beliefs' role in promoting misperceptions and conspiracist ideation. *PLOS ONE*, *12*(9), e0184733. https://doi.org/10.1371/journal.pone.0184733

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying Online Social Networks. *Journal of Computer-Mediated Communication*, *3*(1). https://doi.org/10.1111/j.1083-6101.1997.tb00062.x

Geiß, S., Magin, M., Jürgens, P., & Stark, B. (2021). Loopholes in the Echo Chambers: How the Echo Chamber Metaphor Oversimplifies the Effects of Information Gateways on Opinion Expression. *Digital Journalism*, *9*(5), 660–686. https://doi.org/10.1080/21670811.2021.1873811

Geissinger, A., Laurell, C., & Sandström, C. (2020). Digital Disruption beyond Uber and Airbnb—Tracking the long tail of the sharing economy. *Technological Forecasting and Social Change*, *155*, 119323. https://doi.org/10.1016/j.techfore.2018.06.012

Gil de Zúñiga, H., Jung, N., & Valenzuela, S. (2012). Social Media Use for News and Individuals' Social Capital, Civic Engagement and Political Participation. *Journal of Computer-Mediated Communication*, *17*(3), 319–336. https://doi.org/10.1111/j.1083-6101.2012.01574.x

Goodrow, C. (2021, September 15). *On YouTube's recommendation system*. Blog.Youtube. https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/

Gottfried, J., & Shearer, E. (2017). Americans' online news use is closing in on TV news use. *Pew Research Center*, *7*.

Graham, T. (2015). *Everyday political talk in the internet-based public sphere*. https://doi.org/10.13140/RG.2.1.1217.5524

Gundecha, P., & Liu, H. (2012). Mining Social Media: A Brief Introduction. In *2012 TutORials in Operations Research* (S. 1–17). INFORMS. https://doi.org/10.1287/educ.1120.0105

Haim, M., Graefe, A., & Brosius, H.-B. (2018). Burst of the Filter Bubble?: Effects of personalization on the diversity of *Google News*. *Digital Journalism*, *6*(3), 330–343. https://doi.org/10.1080/21670811.2017.1338145

Hampton, K. N., Rainie, H., Lu, W., Dwyer, M., Shin, I., & Purcell, K. (2014). *Social media and the'spiral of silence'*. PewResearchCenter Washington, DC, USA.

Himelboim, I., Gleave, E., & Smith, M. (2009). Discussion catalysts in online political discussions: Content importers and conversation starters. *Journal of Computer-Mediated Communication*, *14*(4), 771–789. https://doi.org/10.1111/j.1083-6101.2009.01470.x

Hogan, B., Fielding, N., Lee, R., & others. (2008). Analyzing social networks. *The Sage handbook of online research methods*, 141–160.

Huckfeldt, R. R., Johnson, P. E., & Sprague, J. D. (2004). *Political disagreement: The survival of diverse opinions within communication networks*. Cambridge University Press.

Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1–27. https://doi.org/10.1145/3392854

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kaiser, J., & Rauchfleisch, A. (2020). Birds of a Feather Get Recommended Together: Algorithmic Homophily in YouTube's Channel Recommendations in the United States and Germany. *Social Media + Society*, *6*(4), 205630512096991. https://doi.org/10.1177/2056305120969914

Kaiser, J., Rauchfleisch, A., & Córdova, Y. (2021). Comparative Approaches to Mis/Disinformation| Fighting Zika With Honey: An Analysis of YouTube's Video Recommendations on Brazilian YouTube. *International Journal of Communication*, *15*(0). https://ijoc.org/index.php/ijoc/article/view/14802

Kim, M. (2018). How does Facebook news use lead to actions in South Korea? The role of Facebook discussion network heterogeneity, political interest, and conflict avoidance in predicting political participation. *Telematics and Informatics*, *35*(5), 1373–1381. https://doi.org/10.1016/j.tele.2018.03.007

Kirchner, J., & Reuter, C. (2020). Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–27. https://doi.org/10.1145/3415211

Knobloch-Westerwick, S. (2014). *Choice and preference in media use: Advances in selective exposure theory and research*. Routledge.

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., & Baddour, K. (2020). Coronavirus Goes Viral:

Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus*, *12*(3), Article 3. https://doi.org/10.7759/cureus.7255

Krackhardt, D., & Stern, R. N. (1988). Informal Networks and Organizational Crises: An Experimental Simulation. *Social Psychology Quarterly*, *51*(2), 123–140. https://doi.org/10.2307/2786835

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed). SAGE.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Kušen, E., & Strembeck, M. (2018). Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. *Online Social Networks and Media*, *5*, 37–50. https://doi.org/10.1016/j.osnem.2017.12.002

Lange, P. G. (2007). Publicly Private and Privately Public: Social Networking on YouTube. *Journal of Computer-Mediated Communication*, *13*(1), 361–380. https://doi.org/10.1111/j.1083-6101.2007.00400.x

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The people's choice.* (S. vii, 178). Duell, Sloan & Pearce.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062. https://doi.org/10.1126/science.aaz8170

Lee, J. K., Choi, J., Kim, C., & Kim, Y. (2014). Social Media, Network Heterogeneity, and Opinion Polarization. *Journal of Communication*, *64*(4), 702–722. https://doi.org/10.1111/jcom.12077

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting & Electronic Media*, *57*(1), 34–52. https://doi.org/10.1080/08838151.2012.761702

Liu, P., Shivaram, K., Culotta, A., Shapiro, M. A., & Bilgic, M. (2021). The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms. *Proceedings of the Web Conference 2021*, 3791–3801. https://doi.org/10.1145/3442381.3450113

Lu, Y., & Lee, J. K. (2019). Stumbling upon the other side: Incidental learning of counter-attitudinal political information on Facebook. *New Media & Society*, *21*(1), 248–265. https://doi.org/10.1177/1461444818793421

Matthes, J., Rios Morrison, K., & Schemer, C. (2010). A Spiral of Silence for Some: Attitude Certainty and the Expression of Political Minority Opinions. *Communication Research*, *37*(6), 774–800. https://doi.org/10.1177/0093650210362685

McNair, B. (2017). *AN Introduction to Political Communication: Sixth Edition* (6. Aufl.). Routledge. https://doi.org/10.4324/9781315750293

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Melki, J., Tamim, H., Hadid, D., Makki, M., El Amine, J., & Hitti, E. (2021). Mitigating infodemics: The relationship between news exposure and trust and belief in COVID-19 fake news and social media spreading. *PLOS ONE*, *16*(6), e0252830. https://doi.org/10.1371/journal.pone.0252830

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.

Milbradt, B., & Hohnstein, S. (2017). *Mit digitalen Medien groß werden. Wie Smartphone, Tablet und Laptop das Aufwachsen verändern*. *3*, 40.

Mirbabaie, M., Bunker, D., Stieglitz, S., Marx, J., & Ehnis, C. (2020). Social media in times of crisis: Learning from Hurricane Harvey for the coronavirus disease 2019 pandemic response. *Journal of Information Technology*, 026839622092925. https://doi.org/10.1177/0268396220929258

Moessner, M., Feldhege, J., Wolf, M., & Bauer, S. (2018). Analyzing big data in social media: Text and network analyses of an eating disorder forum. *International Journal of Eating Disorders*, *51*(7), 656–667. https://doi.org/10.1002/eat.22878

Munger, K., & Phillips, J. (2022). Right-Wing YouTube: A Supply and Demand Perspective. *The International Journal of Press/Politics*, *27*(1), 186–219. https://doi.org/10.1177/1940161220964767

Nazir, A., Raza, S., & Chuah, C.-N. (2008). Unveiling facebook: A measurement study of social network based applications. *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement Conference - IMC '08*, 43. https://doi.org/10.1145/1452520.1452527

Neubaum, G. (2021). "It's Going to be Out There For a Long Time": The Influence of Message Persistence on Users' Political Opinion Expression in Social Media. *Communication Research*, 009365022199531. https://doi.org/10.1177/0093650221995314

Neubaum, G., & Krämer, N. C. (2017). Monitoring the Opinion of the Crowd: Psychological Mechanisms Underlying Public Opinion Perceptions on Social Media. *Media Psychology*, *20*(3), 502–531. https://doi.org/10.1080/15213269.2016.1211539

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C., & Nielsen, R. (2021). Reuters Institute Digital News Report 2021. *Reuters Institute for Study of Journalism*, *10*, 163.

Noelle-Neumann, E. (1974). The Spiral of Silence A Theory of Public Opinion. *Journal of Communication*, *24*(2), 43–51. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2013). The Extreme Right Filter Bubble. *arXiv:1308.6149 [physics]*. http://arxiv.org/abs/1308.6149

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review*, *33*(4), 459–478. https://doi.org/10.1177/0894439314555329

Oueslati, W., Arrami, S., Dhouioui, Z., & Massaabi, M. (2021). Opinion leaders' detection in dynamic social networks. *Concurrency and Computation: Practice and Experience*, *33*(1). https://doi.org/10.1002/cpe.5692

Ozturk, R., & Coban, S. (2019). Political marketing, word of mouth communication and voter behaviours interaction. *Business and Economics Research Journal*, *10*(1), 245–258.

Pariser, E. (2012). *The filter bubble: What the Internet is hiding from you*. Penguin Books.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Phadke, S., Samory, M., & Mitra, T. (2021). What Makes People Join Conspiracy Communities?: Role of Social Factors in Conspiracy Engagement. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1–30. https://doi.org/10.1145/3432922

Powers, E. (2017). My News Feed is Filtered?: Awareness of news personalization among college students. *Digital Journalism*, *5*(10), 1315–1335. https://doi.org/10.1080/21670811.2017.1286943

Prior, M. (2007). *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.

Pustejovsky, J., & Stubbs, A. (2013). *Natural language annotation for machine learning*. O'Reilly Media.

Qiu, L., Chan, S. H. M., & Chan, D. (2018). Big data in social and psychological science: Theoretical and methodological issues. *Journal of Computational Social Science*, *1*(1), 59–66. https://doi.org/10.1007/s42001-017-0013-6

Radford, J., & Joseph, K. (2020). Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science. *Frontiers in Big Data*, *3*, 18. https://doi.org/10.3389/fdata.2020.00018

Railsback, S. F. (2019). *Agent-based and individual-based modeling: A practical introduction* (2nd edition). Princeton University Press.

Reihanian, A., Minaei-Bidgoli, B., & Alizadeh, H. (2016). Topic-oriented community detection of rating-based social networks. *Journal of King Saud University - Computer and Information Sciences*, *28*(3), 303–310. https://doi.org/10.1016/j.jksuci.2015.07.001

Rodrigues, F. A., & Da F. Costa, L. (2005). SURVIVING OPINIONS IN SZNAJD MODELS ON COMPLEX NETWORKS. *International Journal of Modern Physics C*, *16*(11), 1785–1792. https://doi.org/10.1142/S0129183105008278

Romascanu, A., Ker, H., Sieber, R., Greenidge, S., Lumley, S., Bush, D., Morgan, S., Zhao, R., & Brunila, M. (2020). Using deep learning and social network analysis to understand and manage extreme flooding. *Journal of Contingencies and Crisis Management*, *28*(3), 251–261. https://doi.org/10.1111/1468-5973.12311

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, *28*(4), 394–412. https://doi.org/10.1080/0960085X.2018.1560920

Rotman, D., Golbeck, J., & Preece, J. (2009). The community is where the rapport is—On sense and structure in the youtube community. *Proceedings of the Fourth International Conference on Communities and Technologies - C&T '09*, 41. https://doi.org/10.1145/1556460.1556467

Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, *118*(50), e2102141118. https://doi.org/10.1073/pnas.2102141118

Scheufele, D. A., & Nisbet, M. C. (2013). *Commentary*: Online News and the Demise of Political Disagreement. *Annals of the International Communication Association*, *36*(1), 45–53. https://doi.org/10.1080/23808985.2013.11679125

Schneider, I. K., & Schwarz, N. (2017). Mixed feelings: The case of ambivalence. *Current Opinion in Behavioral Sciences*, *15*, 39–45. https://doi.org/10.1016/j.cobeha.2017.05.012

Scott, J., & Carrington, P. J. (2011). *The SAGE handbook of social network analysis*. SAGE publications.

Shah, D. V., McLeod, D. M., Rojas, H., Cho, J., Wagner, M. W., & Friedland, L. A. (2017). Revising the Communication Mediation Model for a New Political Communication Ecology: Communication Mediation Model. *Human Communication Research*, *43*(4), 491–504. https://doi.org/10.1111/hcre.12115

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2020). An Exploratory Study of COVID-19 Misinformation on Twitter. *arXiv:2005.05710 [cs]*. http://arxiv.org/abs/2005.05710

Shearer, E., & Gottfried, J. (2017). *News use across social media platforms 2017. Pew Research Center; 2017*. https://www.pewresearch.org/journalism/2017/09/07/news-use-across-social-media-platforms-2017/

Smith, A., Anderson, M., & others. (2018). Social media use in 2018. *Pew research center*, *1*, 1–4.

Smith, N., & Graham, T. (2019). Mapping the anti-vaccination movement on Facebook. *Information, Communication & Society*, *22*(9), 1310–1327.

Sohn, D. (2019). Spiral of Silence in the Social Media Era: A Simulation Approach to the Interplay Between Social Networks and Mass Media. *Communication Research*, 009365021985651. https://doi.org/10.1177/0093650219856510

Squazzoni, F., Jager, W., & Edmonds, B. (2014). Social simulation in the social sciences: A brief overview. *Social Science Computer Review*, *32*(3), 279–294.

Stark, B., Stegmann, D., Magin, M., & Jürgens, P. (2020). Are algorithms a threat to democracy. *The rise of intermediaries: a challenge for public discourse. Algorithms Watch Governing Platforms Report*.

Stern, S., & Livan, G. (2021). The impact of noise and topology on opinion dynamics in social networks. *Royal Society Open Science*, *8*(4). https://doi.org/10.1098/rsos.201943

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social Media Analytics: An Interdisciplinary Approach and Its Implications for Information Systems. *Business & Information Systems Engineering*, *Forthcoming*. https://doi.org/10.1007/s11576-014-0407-5

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168.

Stier, S., Bleier, A., Lietz, H., & Strohmaier, M. (2018). Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter. *Political Communication*, *35*(1), 50–74. https://doi.org/10.1080/10584609.2017.1334728

Stöcker, C., & Preuss, M. (2020). Riding the Wave of Misclassification: How We End up with Extreme YouTube Content. In G. Meiselwitz (Hrsg.), *Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis* (S. 359–375). Springer International Publishing.

Su, Y. (2021). It doesn't take a village to fall for misinformation: Social media use, discussion heterogeneity preference, worry of the virus, faith in scientists, and COVID-19-related misinformation beliefs. *Telematics and Informatics*, *58*, 101547. https://doi.org/10.1016/j.tele.2020.101547

Sunstein, C. R. (2007). *Republic.Com 2.0*. Princeton University Press.

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

Theocharis, Y., & Jungherr, A. (2021). Computational Social Science and the Study of Political Communication. *Political Communication*, *38*(1–2), 1–22. https://doi.org/10.1080/10584609.2020.1833121

Thompson, M. M., Zanna, M. P., & Griffin, D. W. (1995). Let's not be indifferent about (attitudinal) ambivalence. *Attitude strength: Antecedents and consequences*, *4*, 361–386.

Treen, K. M. d'I., Williams, H. T. P., & O'Neill, S. J. (2020). Online misinformation about climate change. *WIREs Climate Change*, *11*(5). https://doi.org/10.1002/wcc.665

Turcotte, J., York, C., Irving, J., Scholl, R. M., & Pingree, R. J. (2015). News Recommendations from Social Media Opinion Leaders: Effects on Media Trust and Information Seeking. *Journal of Computer-Mediated Communication*, *20*(5), 520–535. https://doi.org/10.1111/jcc4.12127

Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J. T., Nagler, J., & Tucker, J. A. (2015). Political Expression and Action on Social Media: Exploring the Relationship Between Lower- and Higher-Threshold Political Activities Among Twitter Users in Italy. *Journal of Computer-Mediated Communication*, *20*(2), 221–239. https://doi.org/10.1111/jcc4.12108

Vaccari, C., Valeriani, A., Barberá, P., Jost, J. T., Nagler, J., & Tucker, J. A. (2016). Of echo chambers and contrarian clubs: Exposure to political disagreement among German and Italian users of Twitter. *Social Media+ Society*, *2*(3), 2056305116664221.

Valenzuela, S., Correa, T., & Gil de Zúñiga, H. (2018). Ties, Likes, and Tweets: Using Strong and Weak Ties to Explain Differences in Protest Participation Across Facebook and Twitter Use. *Political Communication*, *35*(1), 117–134. https://doi.org/10.1080/10584609.2017.1334726

Van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C., Matthes, J., Hopmann, D., Salgado, S., Hubé, N., Stępińska, A., Papathanassopoulos, S., Berganza, R., Legnante, G., Reinemann, C., Sheafer, T., & Stanyer, J. (2017). Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association*, *41*(1), 3–27. https://doi.org/10.1080/23808985.2017.1288551

van Atteveldt, W., & Peng, T.-Q. (2018). When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

van der Meer, T. G. L. A., & Jin, Y. (2020). Seeking Formula for Misinformation Treatment in Public Health Crises: The Effects of Corrective Information Type and Source. *Health Communication*, *35*(5), 560–575. https://doi.org/10.1080/10410236.2019.1573295

van Eck, C. W., Mulder, B. C., & van der Linden, S. (2021). Echo Chamber Effects in the Climate Change Blogosphere. *Environmental Communication*, *15*(2), 145–152. https://doi.org/10.1080/17524032.2020.1861048

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Vuong, Q.-H., Nguyen, H. T. T., Pham, T.-H., Ho, M.-T., & Nguyen, M.-H. (2021). Assessing the ideological homogeneity in entrepreneurial finance research by highly cited publications. *Humanities and Social Sciences Communications*, *8*(1), 110. https://doi.org/10.1057/s41599-021-00788-9

Waldherr, A., Hilbert, M., & González-Bailón, S. (2021). Worlds of Agents: Prospects of Agent-Based Modeling for Communication Research. *Communication Methods and Measures*, *15*(4), 243–254. https://doi.org/10.1080/19312458.2021.1986478

Waldherr, A., & Wettstein, M. (2019). Computational Communication Science| Bridging the Gaps: Using Agent-Based Modeling to Reconcile Data and Theory in Computational Communication Science. *International Journal of Communication*, *13*(0). https://ijoc.org/index.php/ijoc/article/view/10588

Walter, S., & Brüggemann, M. (2018). Opportunity makes opinion leaders: Analyzing the role of first-hand information in opinion leadership in social media networks. *Information, Communication & Society*, *23*(2), 267–287. https://doi.org/10.1080/1369118X.2018.1500622

Wattenhofer, M., Wattenhofer, R., & Zhu, Z. (2012). The YouTube Social Network. *Proceedings of the International AAAI Conference on Web and Social Media*, *6*(1). https://ojs.aaai.org/index.php/ICWSM/article/view/14243

Weeks, B. E., Ardèvol-Abreu, A., & Gil de Zúñiga, H. (2015). Online Influence? Social Media Use, Opinion Leadership, and Political Persuasion. *International Journal of Public Opinion Research*, edv050. https://doi.org/10.1093/ijpor/edv050

Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1565

Wilensky, U., & Rand, W. (2015). *An introduction to agent-based modeling: Modeling natural, social, and engineered complex systems with NetLogo*. Mit Press.

Winter, S., & Neubaum, G. (2016). Examining Characteristics of Opinion Leaders in Social Media: A Motivational Approach. *Social Media + Society*, *2*(3), 205630511666585. https://doi.org/10.1177/2056305116665858

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication*, *38*(1–2), 98–139. https://doi.org/10.1080/10584609.2020.1785067

Zeitzoff, T. (2017). How Social Media Is Changing Conflict. *Journal of Conflict Resolution*, *61*(9), 1970–1991. https://doi.org/10.1177/0022002717721392

Zhang, B., Bai, Y., Zhang, Q., Lian, J., & Li, M. (2020). An Opinion-Leader Mining Method in Social Networks With a Phased-Clustering Perspective. *IEEE Access*, *8*, 31539–31550. https://doi.org/10.1109/ACCESS.2020.2972997

Zhang, J., Wang, W., Xia, F., Lin, Y.-R., & Tong, H. (2020). Data-Driven Computational Social Science: A Survey. *Big Data Research*, *21*, 100145. https://doi.org/10.1016/j.bdr.2020.100145

Zillmann, D., & Bryant, J. (1985). *Selective exposure to communication*. Hillsdale, NJ: Lawrence Erlbaum.

Zuiderveen Borgesius, F., Trilling, D., Möller, J., Bodó, B., De Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review. Journal on Internet Regulation*, *5*(1).

Appendix

Research Paper 1: "Identifying Political Sentiments on YouTube: A Systematic Comparison Regarding the Accuracy of Recurrent Neural Network and Machine Learning Models"

Type	Conference
Rights and permission	Reproduced with permission from Springer Nature Open access
Authors	**Röchert, Daniel**; Neubaum, German; Stieglitz, Stefan
Year	2020
Outlet	Multidisciplinary International Symposium on Disinformation in Open Online Media
Publisher	Springer Nature
Permalink/DOI	https://doi.org/10.1007/978-3-030-61841-4_8
Full citation	Röchert, D., Neubaum, G., & Stieglitz, S. (2020). Identifying Political Sentiments on YouTube: A Systematic Comparison Regarding the Accuracy of Recurrent Neural Network and Machine Learning Models. In *Multidisciplinary International Symposium on Disinformation in Open Online Media* (pp. 107-121). Springer, Cham.

Identifying Political Sentiments on YouTube: A Systematic Comparison Regarding the Accuracy of Recurrent Neural Network and Machine Learning Models

Daniel Röchert(✉), German Neubaum, and Stefan Stieglitz

University of Duisburg-Essen, 47057 Duisburg, Germany
daniel.roechert@uni-due.de

Abstract. Since social media have increasingly become forums to exchange personal opinions, more and more approaches have been suggested to analyze those sentiments automatically. Neural networks and traditional machine learning methods allow individual adaption by training the data, tailoring the algorithm to the particular topic that is discussed. Still, a great number of methodological combinations involving algorithms (e.g., recurrent neural networks (RNN)), techniques (e.g., word2vec), and methods (e.g., Skip-Gram) are possible. This work offers a systematic comparison of sentiment analytical approaches using different word embeddings with RNN architectures and traditional machine learning techniques. Using German comments of controversial political discussions on YouTube, this study uses metrics such as F1-score, precision and recall to compare the quality of performance of different approaches. First results show that deep neural networks outperform multiclass prediction with small datasets in contrast to traditional machine learning models with word embeddings.

Keywords: Deep learning · Machine learning · Text classification · Word embeddings · Computational science

1 Introduction

On social media platforms such as YouTube, Facebook, or Twitter, a mass of people interact with each other on a daily basis, commenting on media content such as videos and exchanging their viewpoints on different issues. Since politically and civically relevant communication is becoming more and more prevalent on social media, identifying opinion climates and optimizing approaches remains as an important task for research. To identify the most appropriate method that detects sentiments in political discussions is of pivotal relevance when it comes to grasp dysfunctional communication processes online. For instance, knowing how different opinions are related to each other contributes to assess to what

M. van Duijn et al. (Eds.): MISDOOM 2020, LNCS 12259, pp. 107–121, 2020.
https://doi.org/10.1007/978-3-030-61841-4_8

extent politically homogeneous/heterogeneous cocoons exist. Besides this, identifying sentiments among social media users could also help to asses the opinion climate toward misinformation and to examine that dynamics such misinformation can induce in certain networks. Cross-user generated content such as comments, likes, dislikes or related videos are exchanges of information on a specific topic, also in multi-language context and contain many additional metadata that can be used to analyze user behavior and their current sentiment of a specific topic. Sentiment analysis (SA) also known as opinion mining as a particular form of natural language processing (NLP) is a common tool to grasp communication patterns on social media and is becoming progressively relevant in the research area of social media analytics [34]. Challenges in the area of NLP refer to understanding and processing human communication by machines, not by fixed rules or dictionaries, but rather by training them to learn these complex natural languages. The utilization of SA has become an important method in various domains: product reviews, movie reviews, election campaigns, stock market prediction and social media behavior analysis. The usage of SAs in social media might be used for the decision-making process of companies in order to trace more accurate product strategies based on the customers' current opinions. The more precise the outcome of the SA with regard to product or service reviews, the more effectively strategies can be deployed to prevent crises or to adapt customer requirements. Employing a machine learning approach, a recent study estimated that approx. 60–80% of YouTube comments contain opinions [31]. This makes it highly attractive to identify opinion climates with SA techniques investigating not only public opinion on political issues but also brands and products. Given these numbers, the present study relies on user comments gathered on the platform YouTube. We chose YouTube as a communication platform for our study due to the given prevalence of opinion expressions and its worldwide popularity. The present work applied a comparison of deep learning (subset of machine learning) and traditional machine learning techniques for the categorical classification task to predict the user sentiment score in political YouTube comments and their replies with own input weights of pre-trained word embeddings. Artificial neural models have successfully established themselves in other text classification tasks and achieved good results [26]. However, studies systematically comparing different sentiment analytical combinations are still scarce, especially on the social media platform YouTube with German YouTube comments. With this work, we aim to fill this gap and offer one of the first analyses of German comments on YouTube by using different machine learning techniques. Moreover, this study examines which of those techniques provides better results by combining them with recurrent neural networks and machine learning models. To formalize the overall goals of this paper, the following questions are guiding this research:

RQ1. What is the difference in performance of sentiment classification between recurrent neural networks and traditional machine learning methods?

RQ2. Which of the generated word embedding techniques yields the most accurate results for classification and are any differences detectable among these techniques?

First we crawled YouTube data through the YouTube API, pre-processed the comments and replies by removing inconsistent data and transform them into sentences. Afterwards we transformed all of these sentences into one high dimensional vector also called word embedding which amplifies a dense distributed representation for each word in a high dimension space with the frameworks word2vec and fastText by applying two different techniques such as Skip-Gram and Continuous Bag of Words (CBOW). These word embeddings learned the semantic of their surrounding words and will help to train the deep neural network model as well as the machine learning models.

2 Theoretical Background

2.1 Related Work

Social media have become important communication channels for public interactions in today's digital society. Especially, the investigation of political communication on social platforms, which are examined by means of user-generated comments, plays an increasingly important role in different research areas such as hate speech, misinformation, or political homogenization and polarization. These areas are particularly concerned with the dark side of social media and the ever-growing threat to democracy in society [35]. In particular, the topic of hate speech in social media has generated a lot of attention worldwide in the last few years and is still a current problem for service providers. A study analyzed user comments on the refugee crisis in Germany in 2015/2016 on various news portals [15]. In the study, a binary classifier has been trained using logistic regression, which has achieved a F1-score of 0.67. Further, the researchers have been able to show that many hate words refer to political topics. In addition to the mono-linguistic identification of hate speech, there have been attempts to identify hate speech in different languages using deep-learning techniques and compare them to traditional machine-learning methods [23]. Another aspect that relates to the political context of social media is that these platforms are more often portrayed as a threat to democracy, as they allow interactions between like-minded people. A recent study has examined the YouTube discussion network of comments using opinion-based homogeneity to identify the climate of opinion [29]. The results of the study show that YouTube users reply less on political comments that reflects their own position than on comments that reflect a different opinion. A further problem with social media is that it allows any person to spread claims without any fact-checking. Previous research has shown that the use of social media can increase the impact of fake news and that the main purpose of social media is to influence public opinion as well as political events [18]. In order to prevent this spread of misinformation, several studies focus on

the dissemination and detection of misinformation. A recent study has investigated the identification of fake news from text and images on Twitter using convolutional neural networks and recurrent neural networks [1]. The recurrent neural network has achieved the best performance, thereby making it possible to identify relevant features that are classified as fake news. These "hot topics" in computational research indicate that there is a need to identify those analytical approaches that yield the best classification of sentiments within the large amount of communication data in social media.

2.2 Recurrent Neural Network

Recurrent Neural Networks (RNN) [30] are used for processing sequential information such as language modeling, machine translation, time series prediction or image captioning. The general idea of RNNs is to create a kind of "memory" by performing the same operations on every input values in a feedback connection. This process allows to remember the network from previous processed information by sharing the same weights (parameter sharing) across several time steps in the hidden state and perform the output which depends on the passed information to next network [7]. Especially in NLP, this feature is quite helpful to process sequence of sentences because they mainly follow the same rules across the sequence. Parameter sharing makes it possible to perform the same task at each time-step with different input sequences of variable length and makes it therefore more powerful and dynamic compared to normal feed forward neural networks. It reduces the total number of parameters, which means the RNN does not have to learn the same rules of sequences again and already knows their weights. The formula for processing of sequences of a vector x at every time step looks as follow:

$$h_t = fW(h_{t-1}, x_t) \tag{1}$$

where the activation function f will depend on weights W, which accepts the previous hidden state h_{t-1} as well as the input at the current state x_t. This output will the updated hidden state called h_t.

2.3 Word Embeddings

In 1954, Zellig Harris established the hypothesis that the difference in meaning correlates with the difference in distribution, also known as the distributional hypothesis [10]. This hypothesis is grounded by distributional semantics, which is an active area of research in natural language processing to develop new techniques to capture various semantic phenomena, by computing semantic similarities between words based on their distributional properties in the corpus. One of these techniques is called word embedding and describes the mapping process of words from a vocabulary into a high dimensional vector spaces by keeping semantically related words close together. It uses an embedding matrix $E \in \mathbb{R}^{|V| \times d_w}$ where d_w is the dimensionality of the embedding space and $|V|$ is the size of the vocabulary. In previous research, this technique is an efficient

way to improve and simplify many NLP applications such as machine translation [20,39], spelling correction [13] or SA [4,17]. In the context of a SA with classification problem, word embeddings are mainly used to include the semantic connections of words in the analysis to develop better and more accurate predictions. A widely used unsupervised word embedding algorithm is called word2vec[1], which has been developed by Mikolov et al. from Google and contains a two-layer neural network, which uses text data as input and transforms the output as a set of high dimensional vectors [19]. Another unsupervised distribution semantic model is called fastText[2] which has been developed by Facebook and is essentially an extension of word2vec model. The main difference of both methods is that the fastText algorithm supports the use of n-grams, which improve the syntatics tasks by taking morphological information into account [3]. Both models have implemented the CBOW and the Skip-Gram methods for computing vector representations of words and are based on hierarchical softmax and negative sampling. The Skip-Gram method has been introduced by Mikolov et al. and predicts potential neighboring words based on a target word [19]. Whereas the CBOW technique uses the context of the neighboring words and predicts the target word. Negative sampling is a modification of an approach called Noise Contrastive Estimation (NCE) [8]. The main idea of the sampling-based approach is to reduce the performance of computational by noise contrastive estimation with several negative examples. An experiment has shown that the negative sampling method is the most efficient algorithm independent from the language used [22]. The present work is intended to compare different combinations of techniques (word2vec and fastText) and methods (Skip-Gram and CBOW), generating unique word vectors that represent the projection of YouTube comments in a continuous vector space.

3 Research Method

This section deals with the sentiment analysis by using deep leaning methods such as RNN to analyze two controversial topics in Germany that were discussed on the social media platform YouTube. The following Fig. 1 demonstrates the process structure of the approach. First, we crawled YouTube data through the YouTube API, preprocessed the comments and replies by removing inconsistent data and transform them into sentences. Afterwards, we transformed all of these sentences into one high dimensional vector also called word embedding which amplifies a dense distributed representation for each word in a high dimension space with the frameworks word2vec and fastText by applying two different techniques such as Skip-Gram and CBOW. These word embeddings learned the semantic of their surrounding words and will help to train the model. The recurrent neural network has been used to initialize these embedding weights to train the network and to create a classifier for further predictions.

[1] https://code.google.com/archive/p/word2vec/.
[2] https://fasttext.cc.

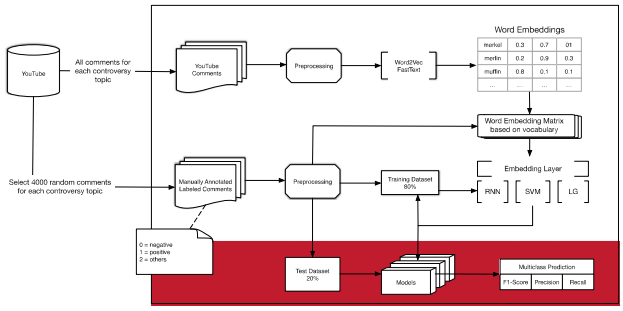

Fig. 1. Process of training and evaluation.

3.1 Data Acquisition

The data of this study were gathered on YouTube crawled by the YouTube API and conducted on May 15th 2018. The data comprised comments and replies of two controversial topics in Germany. The first crawl started with the search criterion: "Kopftuchverbot in Deutschland - Headscarf ban in Germany", for this search we collected 320 unique videos with a maximum number of 48,354 comments and replies. The debate "Wearing religious headscarves" is met with supporters who claim that to fulfil freedom of religion it needs to be allowed while opponents state that headscarves are a symbol for female oppression. For the second query, we used the search terms: "Adoptionsrecht für homosexuelle Paare - Adoption rights for homosexual couples" and contains 15,889 comments and replies with 266 unique videos. In Germany, the debate "Adoption rights for homosexual couples" continues to cause debate. Advocates argue that there is no reason to not allow joint adoption for homosexual partners, whereas opponents argue that every child needs a mother and a father, reflecting their normative ideas of family life. Both topics were selected because they highlight current and controversial issues in society and thus have a lot of potential for discussion in social media. We assume that both controversial topics exhibit a sentiment diversity (i.e. a similar distribution of pros and cons). YouTube labels each video with their own categoryID[3], in this case we use categoryID of 25, which stands for "Politics and News" in the YouTube API. After filtering the comments and replies for both search criterion's, we had a data pool of 14,277 comments and replies for the first dataset "Headscarves" and 8,443 comments and replies for the second dataset "Adoption rights".

[3] https://developers.google.com/youtube/v3/docs/videoCategories.

3.2 Annotator Agreement

We selected two well trained independent annotators who received the same dataset with 4,000 randomly selected comments and replies for each topic. The term "well trained" refers to annotators who have received personal instruction on the topics presented. Furthermore, the explanation of the coding scheme included example sentences and their coding was outlined and clarified. Through the personal instruction, questions and problems could be clarified to eliminate inconsistencies. The data were labelled with one of three classes which were mutually exclusive: *negative*, *positive* and *others* based on an existing coding scheme [29]. While there was 86.6% percent agreement for the topic of headscarf ban, 82.5% of percent agreement was reached for the topic of adoption right for homosexual couples. In order to ensure better results for all machine learning models, we have decided to use only those records that have an equal match between both annotators for later analysis.

3.3 Data Preparation

Besides the labeling process, another complex task is the cleaning of unstructured data. In terms of data preparation, cleaning unstructured data guarantees that algorithms can classify better and compute more accurate results with the pre-prepared data [34]. Therefore, the main workflow of cleaning the text by regular expression includes: removing hyperlinks and usernames; removing special symbols and numerical values; converting words into lowercase and assigning smilies into three different word categories such as "emotionhappy", "emotionsad", "emotionlaugh."

After cleaning the text, it is necessary to split it into sentences or paragraphs, which is required for the word embedding models word2vec and fastText in Python. For the supervised learning problem, it is required to separate the entire dataset into the training and test datasets. Each dataset is split into training set (80%) and test set (20%). This separation has to be randomized to guarantee that there is no noise in the dataset. We used 5-fold cross-validation on the training dataset to evaluate the performance of all models with a fixed combination of manual-based hyperparameters.

3.4 Unsupervised Learning

In our approach, we used a Python implementation of the word2vec and fastText from Gensim, which is used for NLP task like topic modeling, document indexing and similarity retrieval [27]. We decided to generate our own word embeddings because recent studies have shown that the creation of domain-specific word embeddings such as (crisis, patent) in particular can enhance the performance of the classification, compared to the pre-trained embeddings of Wikipedia or Google News, which are more suitable for more general classification tasks [16, 28]. We created for each method 300 high dimensional word embeddings on basis of the comments and replies of the whole corpus where the words represent as

unigrams. As mentioned earlier, both models word2vec and fastText apply Skip-Gram and CBOW techniques and use the same parameter settings to make them comparable afterwards in the evaluation of the sentiment model. The applied parameters with a brief description of their functionalities are presented in the following:

1. *size: represents the dimension of the feature vectors.*
2. *min_count: represents the minimum frequency per token to filter rare words.*
3. *alpha: represents the learning rate of the network.*
4. *iter: represents the epoch over the corpus to update the weights.*
5. *sample: represents the threshold for configuring which higher-frequency words are randomly downsampled.*
6. *negative: represents the amount of how many "noise words" should be included during training.*

The final parameters that have been used for all word embeddings are *size* with a value of 300, *min_count* with a value of 5, *alpha* with a value of 0.01, *iter* with a value of 15, *sample* with a value of 0.05, and *negative* with a value of 15. Since our vocabulary has a size of 11,435 and the dataset contains 126,362 clean sentences with 1,939,663 tokens, we have deliberately opted for a larger dimension (300). For the further process, we chose negative sampling as baseline in this work, which can improve the computation of word embeddings for frequent words and also decrease the performance of training speed of the neural network [21]. Table 1 demonstrates the representation of the embedding matrix to find the top four most similar entities for the word "kopftuch" (headscarf) and "homosexuell" (homosexual). Looking at the results of the two methods, it is noticeable that the most similar words of word2vec are rather different, but nevertheless relevant to the context. On the other hand, the entities of fastText consist of many variations that are very close to the actual word. It is therefore relevant to examine to what extent which of the two methods delivers the better results in the prediction.

3.5 Supervised Learning

For the supervised learning task, we implemented our model based on Keras with a TensorFlow backend [5]. Keras is a Python library for developing deep neural networks. The baseline models have been implemented as well in Python, but with the sci-kit library for machine learning in Python [24]. The implementation and configuration of all recurrent neural networks share all the same parameters to make them comparable with the different combination of word embeddings. We used a many-to-one model for our architecture, where the input of the network is characterized by sentences with variably sized and multiple words. The first layer of our sequential model is the embedding layer initialized by a dimension 300 and an input length of 100. After this layer, we set a recurrent layer with 64 hidden units, an internal dropout rate of 0.1 and a recurrent dropout of 0.1. The main reason for the regularization of a neural network is the increase

Table 1. Word similarities of the words “kopftuch” and “homosexuell”.

	word2vec	
	CBOW	**SkipGram**
Target word: kopftuch	hijab (0.68)	tragen (0.66)
	koptuch (0.66)	hijab (0.60)
	kopftücher (0.60)	koptuch (0.59)
	tuch (0.57)	minirock (0.59)
	FastText	
	kopftuchs (0.97)	kopftuchzwang (0.82)
	kopftuchgebot (0.97)	kopftuchfrau (0.82)
	kopftuchzwang (0.96)	kopftuchs (0.81)
	Koftuch (0.94)	kopftuchgebot (0.8)
Target word: homosexuell	**word2vec**	
	schwul (0.80)	schwul (0.62)
	heterosexuell (0.79)	lesbisch (0.60)
	bisexuell (0.73)	heterosexuell (0.60)
	lesbisch (0.72)	bisexuell (0.59)
	FastText	
	homosexuel (0.97)	homosexuel (0.93)
	homosexuele (0.97)	homosexuele (0.93)
	homosexuelle (0.95)	homosexuellen (0.82)
	homosexuelles (0.95)	homosexuelle (0.82)

in performance and its applicability to unseen data beyond the training data and to avoid overfitting. Especially for small datasets, neural networks are more inclined to overfit than on large datasets because they are used to learn from large data. For regularization of our network, we decided to employ to usual methods: applying dropout to the networks [11], using L_2 weight regularization as well as class weights. The general idea of dropout is to avoid co-adaptations by applying random dropout units during the training of the neural network [33]. Using dropout can greatly reduce overfitting in RNNs [37]. Besides the dropout regularization, we used L_2 weight decay to reduce the complexity of the softmax function. Readjustment of the class weights have been applied to re-balance the classes and make them more reasonable and equally considered during the training. Classes that appear in the dataset often achieve low weights, whereas infrequent classes receive higher weights to re-balance the training. As an optimization function to train our network, we choose the extension to stochastic gradient descent called Adam [14] with the categorical cross entropy loss function, suited to multi-class classification problems. The output layer are characterized by three neurons with a softmax activation function for predicting the probability distribution for each class. Further, we trained the model with a

mini-batch size of 10 and set the number of epochs of 100. Because our dataset is small for training and testing we chose a small training batch size, as well as small hidden units of the neural networks to increase the accuracy of the prediction.

3.6 Baseline Models

We have also implemented traditional machine learning models, so that we can identify how neural networks perform in comparison to methods which can perhaps better handle smaller datasets. In a study where Chinese short texts with public financial documents were classified, it was shown that the support vector machines as well as logistic regressions achieved the best results of the performance of the prediction [36]. More precisely, by comparing machine learning models, it was shown that logistic regression in the area of product reviews [25] or BBC news [32] achieved better results than other classical machine learning models such as k-nearest-neighbors or random forest. Apart from logistic regression, there exist also several studies showing that the Support Vector machine was successfully applied for text classification and reached the best performance in multi-class prediction [16,38]. Based on the positive results of the previously stated studies, we have decided to use SVM and logistics regression as baseline machine learning techniques. In general, machine learning models such as support vector machines or logistic regression cannot directly handle word embeddings, which are represented in a high-dimensional space, therefore we have to prepare our 300 dimensional word embeddings into one dimensional by using the average value of each word vectors. This allows us to represent each word by an average value and have been successfully implemented on other studies [2]. The following machine learning techniques have been performed:

- Support Vector Machines (SVM): are based on the margin maximization principle and used for non-linear and linear regression and classification tasks. The SVM uses a penalty parameter C of the value of 100, which is characterized as the error term, the smaller the value, the stronger is the regulation of the model. As well, we applied a *linear* kernel and balanced class weights to the model.
- Logistic Regression (LG): as well as SVM, logistic regression is a supervised learning algorithm to estimates the probability of a categorical dependent variables by computing the sigmoid function. Like for the recurrent neural networks, we avoided overfitting by applying L_2 weights regularization with the *saga* solver. Also we used balanced class weights to adjust the imbalanced distribution of classes.

4 Results

Since the models have been trained successfully, the information of the models can be extracted and used for analysis. The results for the prediction on the test

datasets are shown in Table 2. We used precision, recall and F1-score to measure the performance of three different classes. For multi-class tasks which are imbalanced, it is recommended to apply weighted F1-score, which computes the average for each class. The results for the performance F1-score reveal two main features. First, the results show that in general all recurrent neural networks outperform the machine learning models. The best performance was achieved with RNNs obtained by combining word2vec with CBOW for both datasets. Second, focusing on the different word embedding methods like Skip-Gram and CBOW, the results indicate that CBOW performs better than Skip-Gram, especially for RNNs, but this does not apply to the remaining results. Looking at the results for the individual word embeddings techniques "word2vec" and "fastText", no particular difference is noticeable because the F1-values are generally very similar to each other.

Table 2. Evaluation result of deep learning and traditional machine learning methods on test dataset.

			Adoption rights			Headscarves		
Models	Technique	Method	F1-score	Precision	Recall	F1-score	Precision	Recall
RNN	**word2vec**	Skip-Gram	0.715	0.726	0.706	0.789	0.794	0.784
		CBOW	**0.746**	**0.748**	**0.744**	**0.823**	**0.815**	**0.835**
	fastText	Skip-Gram	0.724	0.739	0.738	0.754	0.806	0.724
		CBOW	0.741	0.731	0.760	0.755	0.793	0.731
SVM	word2vec	Skip-Gram	0.565	0.717	0.509	0.543	0.798	0.470
		CBOW	0.568	0.721	0.512	0.543	0.798	0.470
	fastText	Skip-Gram	0.567	0.719	0.512	0.544	0.801	0.471
		CBOW	0.560	0.723	0.503	0.552	0.798	0.480
LG	word2vec	Skip-Gram	0.597	0.704	0.550	0.647	0.783	0.585
		CBOW	0.592	0.703	0.544	0.642	0.783	0.577
	fastText	Skip-Gram	0.600	0.710	0.553	0.649	0.783	0.587
		CBOW	0.581	0.699	0.532	0.629	0.773	0.562

5 Discussion

This study offered a systematic comparison of combinations consisting of different sentiment analytical approaches such as deep neural networks and machine learning models. With regard to RQ1, we can conclude that our approach has demonstrated that the artificial neural network models outperform usual machine learning models by embedding high dimensional vectors. In order to have a fair comparison, hyperparameters were kept constant in this study. The fact that deep neural networks generally reach higher F1-scores may be explained by different factors: First, the weaker results of machine learning methods can

be explained by the fact that they cannot capture the high-dimensional word vectors during training, but only receive averaged word vectors for all words in the corpus. As a result, important information is no longer provided during the computation and performance deteriorates. Second, deep neural networks might reach a higher level of accuracy when hyperparameters are determined by grid search or random search in accordance with the dataset at hand. Third, it must be noted that the dataset is imbalanced, methods to weight the classes are beneficial, but more effective would be actual datasets with equally distributed classes. Given that the category with the most comments was the *others* class, all methods might benefit more from a dataset that has a larger portion of positive versus negative comments whose context are easier to identify (than from *others* comments). When considering the normalized confusion matrix, the class most frequently predicted in neural networks is "others", which therefore has a positive effect on the F1-score, since this class is most frequently represented in the dataset. Regarding RQ2, it can be concluded that word embeddings have significantly improved the performance of RNN compared to the traditional ML models. Due to the different models, however, it is not possible to determine exactly which method and technique is the best because the different combinations of word embeddings have computed relatively similar outcomes.

6 Further Research

To conclude, the present study revealed that with small datasets of user comments on YouTube, deep neural networks outperform machine learning models. For future work, it would be interesting to improve some features to achieve more precise results in the prediction. The first improvement in analysis might consist of applying advanced models and techniques to compute even more accurate predictions. Simple RNNs are often used for processing long-term sequences like documents, however studies have shown that RNNs are mainly suitable for short term dependencies because of the vanishing gradient or exploding gradient problem [12], which makes them inaccurate for tasks that require long-term sequences. This problem appears when training deep neural networks to learn dependencies by backpropagation through time over long time steps, which can reach extremely high or exponentially small values of gradients. To avoid this kind of problem, it is commendable to apply other recurrent neural network architectures such as long-short term memory. For further research, it would be advisable to implement and compare the improved RNNs like long-short term memory networks as well. Furthermore, it would be a reasonable idea to utilize further machine learning algorithms such as naive bayes or random forest, which are not based on word embeddings but on term frequency times inverse document frequency vectors to extend the systematic comparison and test which combined approaches offer more accurate results. It does not require word embedding but is also used for SA [6,9]. While word embedding links the semantics of sentences, term frequency times inverse document frequency computes the importance of a term inside a comment by their frequency of the entire dataset. In addition, a

further aspect that should be considered when using word embeddings in future research is the comparison with already existing pre-trained models such as Wikipedia or Google News to be able to make semantic comparisons between these and own domain-specific models and to take into account which models are better suited for classification. Another necessary step is to perform SA with other languages in order to achieve greater diversity and compare them against each other. In addition to political and controversial topics, it would also be appropriate to collect data from product reviews or unboxing videos and evaluate the comments with the aid of SA to gain experience in this field as well.

Acknowledgments. This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia (Grant Number: 005-1709-0004), Junior Research Group "Digital Citizenship in Network Technologies" (Project Number: 1706dgn009).

References

1. Ajao, O., Bhowmik, D., Zargari, S.: Fake news identification on twitter with hybrid CNN and RNN models. In: Proceedings of the 9th International Conference on Social Media and Society, pp. 226–230 (2018). https://doi.org/10.1145/3217804.3217917
2. Bayot, R.K., Gonçalves, T.: Author profiling using SVMS and word embedding averages. In: CLEF (Working Notes), pp. 815–823 (2016)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
4. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326 (2015)
5. Chollet, F., et al.: Keras (2015). https://keras.io
6. Ghag, K., Shah, K.: SENTITFIDF - sentiment classification using relative term frequency inverse document frequency. Int. J. Adv. Comput. Sci. Appl. **5**(2) (2014). https://doi.org/10.14569/IJACSA.2014.050206
7. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning, vol. 1. MIT Press, Cambridge (2016)
8. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 9, pp. 297–304. PMLR (2010). http://proceedings.mlr.press/v9/gutmann10a.html
9. Haddi, E., Liu, X., Shi, Y.: The role of text pre-processing in sentiment analysis. Procedia Comput. Sci. **17**, 26–32 (2013). https://doi.org/10.1016/j.procs.2013.05.005
10. Harris, Z.S.: Distributional structure. Word **10**(2–3), 146–162 (1954). https://doi.org/10.1080/00437956.1954.11659520
11. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
12. Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertainty, Fuzz. Knowl.-Based Syst. **6**(02), 107–116 (1998). https://doi.org/10.1142/S0218488598000094

13. Kilicoglu, H., Fiszman, M., Roberts, K., Demner-Fushman, D.: An ensemble method for spelling correction in consumer health questions. In: AMIA Annual Symposium Proceedings. vol. 2015, p. 727. American Medical Informatics Association (2015)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR abs/1412.6980 (2015)
15. Köffer, S., et al.: Discussing the value of automatic hate speech detection in online debates. Multikonferenz Wirtschaftsinformatik (MKWI 2018): Data Driven X-Turning Data in Value, Leuphana, Germany (2018)
16. Li, H., Caragea, D., Li, X., Caragea, C.: Comparison of word embeddings and sentence encodings as generalized representations for crisis tweet classification tasks. en. In: New Zealand p. 13 (2018)
17. Li, Q., Shah, S., Liu, X., Nourbakhsh, A., Fang, R.: Tweetsift: tweet topic classification based on entity knowledge base and topic enhanced word embedding. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 2429–2432. ACM (2016). https://doi.org/10.1145/2983323.2983325
18. Marwick, A., Lewis, R.: Media Manipulation and Disinformation Online. Data & Society Research Institute, New York (2017)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
20. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. NIPS 2013, vol. 2. pp. 3111–3119 (2013)
22. Naili, M., Chaibi, A.H., Ghezala, H.H.B.: Comparative study of word embedding methods in topic segmentation. Procedia Comput. Sci. **112**, 340–349 (2017). https://doi.org/10.1016/j.procs.2017.08.009
23. Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., Yeung, D.Y.: Multilingual and multi-aspect hate speech analysis. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4675–4684. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1474
24. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
25. Pranckevicius, T., Marcinkevicius, V.: Comparison of Naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Balt. J. Mod. Comput. **5** (2017). https://doi.org/10.22364/bjmc.2017.5.2.05
26. Rao, A., Spasojevic, N.: Actionable and political text classification using word embeddings and LSTM. CoRR abs/1607.02501 (2016)
27. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, May 2010. https://doi.org/10.13140/2.1.2393.1847
28. Risch, J., Krestel, R.: Domain-specific word embeddings for patent classification. Data Technol. Appl. (2019). https://doi.org/10.1108/DTA-01-2019-0002

29. Röchert, D., Neubaum, G., Ross, B., Brachten, F., Stieglitz, S.: Opinion-based homogeneity on YouTube. Comput. Commun. Res. **2**(1), 81–108 (2020). https://doi.org/10.5117/CCR2020.1.004.ROCH
30. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533 (1986). https://doi.org/10.1038/323533a0
31. Severyn, A., Moschitti, A., Uryupina, O., Plank, B., Filippova, K.: Multi-lingual opinion mining on YouTube. Inf. Process. Manage. **52**(1), 46–60 (2016). https://doi.org/10.1016/j.ipm.2015.03.002
32. Shah, K., Patel, H., Sanghvi, D., Shah, M.: A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Hum. Res. **5**(1), 1–16 (2020). https://doi.org/10.1007/s41133-020-00032-0
33. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
34. Stieglitz, S., Mirbabaie, M., Ross, B., Neuberger, C.: Social media analytics-challenges in topic discovery, data collection, and data preparation. Int. J. Inf. Manage. **39**, 156–168 (2018). https://doi.org/10.1016/j.ijinfomgt.2017.12.002
35. Sunstein, C.R.: # Republic: Divided Democracy in the Age of Social Media. Princeton University Press, Princeton (2018)
36. Wang, Y., Zhou, Z., Jin, S., Liu, D., Lu, M.: Comparisons and selections of features and classifiers for short text classification. IOP Conf. Ser. Mat. Sci. Eng. **261**, 012018 (2017). https://doi.org/10.1088/1757-899x/261/1/012018
37. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent neural network regularization. arXiv preprint arXiv:1409.2329 (2014)
38. Zhang, M., Ai, X., Hu, Y.: Chinese text classification system on regulatory information based on SVM. IOP Conf. Ser. Earth Environ. Sci. **252**, 022133 (2019). https://doi.org/10.1088/1755-1315/252/2/022133
39. Zou, W.Y., Socher, R., Cer, D., Manning, C.D.: Bilingual word embeddings for phrase-based machine translation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1393–1398 (2013)

Research Paper 2: "Opinion-based homogeneity on YouTube: Combining sentiment and social network analysis"

Type	Journal
Rights and permission	Re-used from Amsterdam University Press Open access
Authors	**Röchert, Daniel**; Neubaum, German; Ross, Björn; Brachten, Florian; Stieglitz, Stefan
Year	2020
Outlet	Computational Communication Research (CCR)
Publisher	Amsterdam University Press
Permalink/DOI	https://doi.org/10.5117/CCR2020.1.004.ROCH
Full citation	Röchert, D., Neubaum, G., Ross, B., Brachten, F., & Stieglitz, S. (2020). Opinion-based homogeneity on YouTube: Combining sentiment and social network analysis. *Computational Communication Research*, *2*(1), 81-108.

Opinion-based Homogeneity on YouTube

Combining Sentiment and Social Network Analysis

Daniel Röchert, German Neubaum, Björn Ross, Florian Brachten, Stefan Stieglitz

CCR 2 (1): 81–108
DOI: 10.5117/CCR2020.1.004.RÖCH

Abstract

When addressing public concerns such as the existence of politically like-minded communication spaces in social media, analyses of complex political discourses are met with increasing methodological challenges to process communication data properly. To address the extent of political like-mindedness in online communication, we argue that it is necessary to focus not only on ideological homogeneity in online environments, but also on the extent to which specific political questions are discussed in a uniform manner. This study proposes an innovative combination of computational methods, including natural language processing and social network analysis, that serves as a model for future research examining the evolution of opinion climates in online networks. Data were gathered on YouTube, enabling the assessment of users' expressed opinions on three political issues (i.e., adoption rights for same-sex couples, headscarf rights, and climate change). Challenging widely held assumptions on discursive homogeneity online, the results provide evidence for a moderate level of connections between dissimilar YouTube comments but few connections between agreeing comments. The findings are discussed in light of current computational communication research and the vigorous debate on the prevalence of like-mindedness in online networks.

Keywords: machine learning, echo chamber, social network analysis, computational science, opinion-based homogeneity

Social media such as YouTube, Facebook, or Twitter have fundamentally changed people's political communication by offering the opportunity to exchange opinions across time and geographical barriers. At the same time, there are risks associated with the use of social media, such as being exposed to manipulative agents like social bots, the viral spread of misinformation, or the formation of echo chambers, i.e., online spaces in which users exclusively encounter information and opinions in line with their own.

According to current research, these risks could (a) undermine the heterogeneity of opinion climates (Graham, 2015), (b) narrow (political) world views and even convey distorted pictures of public opinion to individual users (Neubaum & Krämer, 2017), and (c) foster a polarization of viewpoints and fragmentation of society (Sunstein, 2017). Empirical studies using computational methods (i.e. network analyses) have found that users in networks such as Twitter indeed move in ideologically homogeneous clusters, but are still confronted time and again with information and opinions divergent from their own (e.g., Bakshy, Messing, & Adamic, 2015; Guo, Rohde, & Wu, 2018) which, in turn, has been shown to contribute to depolarization (Beam, Hutchens, & Hmielowski, 2018).

While these examples provide initial evidence on ideological homogeneity in online networks (e.g., are Democrats more likely to be connected to Democrats?), a focus on ideology can only serve as a proxy for the extent to which individuals encounter views that are dissimilar to theirs. When it comes to analyzing the connection between similar and dissimilar stances, it seems more informative to focus on specific politically and civically relevant topics that are factually debated, that is, on the content of the discussion.

Against this backdrop, the present study proposes an analytical approach that addresses specific issue-related discussions on social media and opinion-based homogeneity therein. Accordingly, we refer to opinion-based homogeneity as the extent to which a set of political opinions that are similar are connected with each other (relative to the extent to which they are connected to dissimilar opinions). While ideological homogeneity operates on a general group level in terms of being, for example, liberal or conservative, opinion-based homogeneity requires a reference to specific political topics. This topic-oriented approach is thought to offer a more nuanced view of the nature of homogeneous versus heterogeneous online discussions and the prevalence of like-minded spaces when it comes to political discussions.

To our knowledge, no research has addressed online homophily based on opinion-based homogeneity by combining natural language processing

and social network analyses. Using the amalgamation of these two approaches, this study investigates to what extent citizens' opinion expressions in the form of user-generated comments are related to each other when they represent a similar stance on a politically relevant question. To this end, written German user-generated comments on political issues were analyzed.

Literature in this area has been limited to the investigation of social media platforms such as Facebook and Twitter, largely neglecting the most popular video-sharing platform YouTube. According to the website ranking platform SimilarWeb[1], YouTube is visited more often (28.9 billion visits in the last six months, as of November 2019) than Facebook (24.6 billion), and significantly more than Twitter (4.6) or Instagram (4.1). YouTube is turning more and more into a platform where users not only watch videos, but especially young users form communities to discuss videos or topics, and exchange opinions on current politically relevant debates (YouGov & BRAVO, 2017). Thus, it seems a pressing need to investigate the potential existence of political like-mindedness on the social platform YouTube. To formalize the general objectives of this paper, two questions guide this research:

RQ1. How high is the prevalence of opinion-based homogeneity among YouTube comments on specific political topics?

When addressing online homogeneity, there might still be differences between homogeneity at a large scale, referring to the whole network (e.g., the whole platform) which covers the full range of the topical discussions, and sub-networks in which discussions are based on reciprocal responses. Consequently, we ask:

RQ2. How does opinion-based homogeneity vary between analyses on a macro level (i.e., focusing on discussions across the full network) and a micro level (i.e., focusing on sub-networks) among YouTube comments?

To address these questions, this paper presents a combined approach of social network analysis (SNA) and sentiment analysis (SA). Crawling a multi-content social networking platform such as YouTube allows us to create a model based on unstructured German YouTube comments to predict the sentiment score of multiple users toward specific controversial topics. In particular, the present approach uses support vector machines (SVM) to predict the sentiment score on German comments of controversial political discussions on YouTube. These analyses were run for three different politically relevant topics: the right of same-sex couples to adopt children,

a ban on headscarves, and climate change. These topics have been discussed extensively in the public and represent good examples of divisive issues that are associated with fundamental moral questions.

Background

Political Homogeneity in Online Communication

In many instances, it has been suggested that politically and civically relevant communication on social media can hold individual users captive in spaces in which they are exposed to political views that are in line with their pre-existing opinions (i.e., so-called "echo chambers") (Boutyline & Willer, 2017; Sunstein, 2017). In light of democratic ideals, politically homogeneous spaces are assumed to lead to political polarization and radicalization since users are allegedly caught in self-reinforcing networks which, in the long run, could become more extreme (Prior, 2007). When it comes to analyzing whether and how individuals might get "caught" in those like-minded networks, different (non-mutually exclusive) scenarios are conceivable (Flaxman, Goel, & Rao, 2016; Geschke, Lorenz, & Holtz, 2019): (a) users actively homogenize their network and, therefore, their information sources, (b) algorithms shape the ideological environment of users, or (c) users are incidentally exposed to a thread of like-minded information (e.g., when comments refer to other comments that are uniform in the stance they express). The present work focuses on the latter scenario and investigates to what extent user-generated comments on political questions are related to congenial comments by others.

Initial evidence focusing on political homogeneity online showed that people are indeed connected to like-minded users to a larger extent than to politically opposing users in the United States (e.g., Bond & Messing, 2015; Boutyline & Willer, 2017). Theoretically, this pattern can be explained by the notion of selective exposure (Colleoni, Rozza, & Arvidsson, 2014; Knobloch-Westerwick, 2014; Zillmann & Bryant, 1985): People experience positive emotions when consuming information that conforms to their pre-existing views and feel stressed when the information contradicts their views. As a result, they seek out situations in which they are exposed to information that is in line with their views. This makes them more likely to affiliate with like-minded others and create homogeneous groups. While social media users may commonly be fully in control of their virtual acquaintances (e.g., in terms of friending or following someone or a news channel), they may not have full control over the information and stances

they are exposed to incidentally, for instance, when browsing through certain Facebook or YouTube news channels (Lu & Lee, 2018). Following this logic, it seems worthwhile to ask to what extent users are actually exposed to and in contact with opinions they disagree with.

Empirical research addressing the potential existence of echo chambers in online networks has been based on two different approaches: On the one hand, survey research has relied on subjective estimates by social media users. This line of research, asking participants how frequently they are exposed to opinion or ideological diversity, has shown that on social media, people are incidentally exposed to heterogeneous opinions (e.g., Kim, 2018; Lee, Choi, Kim, & Kim, 2014; Lu & Lee, 2018; Vaccari et al., 2016). On the other hand, another series of studies used observational data and made use of computational methods, especially focusing on SNAs.

The Assessment of Political Homogeneity Online based on Social Network Analyses

As a widely used method in research focusing on political homogeneity, SNA examines the properties of social networks – networks composed of people and their social connections with one another. In SNA, the property of an individual to seek social connections to other individuals with similar characteristics is called homophily. In other words, homophily is described as "the principle that a contact between similar people occurs at a higher rate than among dissimilar people" (McPherson, Smith-Lovin, & Cook, 2001, p. 416).

Several studies have therefore used network analysis to examine political like-mindedness in network data. Bakshy et al. (2015) analyzed Facebook data to examine the political homogeneity of friend networks to identify whether users read and share messages that are more consistent with their political ideological beliefs than cross-cutting content. Their findings showed that about 20% of users' Facebook friends were from the opposing party, which increases the probability that users will receive content that diverges from their own ideology. Another study focused on Twitter data to determine ideological homogeneity by analyzing 3.8 million Twitter users and a dataset of almost 150 million tweets on political and non-political topics (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015). Their results revealed that Democrats were significantly more likely than Republicans to be involved in the cross-ideological dissemination of political and non-political information. Recently, Del Valle and Bravo (2018) ran an SNA of the Twitter network among Catalan parliamentarians and how information flows among them. Their study found that representatives are more likely to interact with members of their own party who share the same political interests.

On a methodical level, to identify echo chambers which are characterized by "disproportionate connections among ideologically similar political communicators" (Jasny, Waggle, & Fisher, 2015), the structural properties of the social network – specifically, the likelihood of connections between members of a group – need to be compared with the political views of the members of the network. If the two are related, that is, if there is a group of individuals with a disproportionately high density of intra-group connections compared to the number of outside connections, whose members share political views that they do not share with non-members, this group can be considered an echo chamber, that is, a politically homogeneous communication space.

The identification of political homogeneity, thus, requires two steps: identifying a group (i.e., a subset) of users who agree politically, and measuring whether there is a disproportionately high number of connections between group members. Researchers have used various methods for both steps.

A useful way of quantifying the relationship between intra-group and inter-group connections is the E-I index. It was presented for the first time by Krackhardt and Stern (1988) and compares the strength of internal connections between members of a class to the strength of external connections to non-members. Other studies examining political echo chambers used similar methods and based their conclusions on the E-I index, e.g., to assess the fragmentation between pairs of discussion networks or the effect of tie strength on the polarization in such networks (Bright, 2018; Chan & Fu, 2017). By using the E-I index, it is possible to quantify the degree to which members of a group interact with each other, as opposed to interacting with others outside the group.

In addition, the identification of political homogeneity requires information about the political affiliations or views of the members of a social network. With observational network data at hand, the arguably most accurate source for inferring an actor's political views is the set of posts and comments in which he/she expressed his/her viewpoints. Working with unstructured text data poses especially difficult challenges (Stieglitz, Mirbabaie, Ross, & Neuberger, 2018), but there are a few studies that have used methods from natural language processing to tackle this problem.

Identifying Political Opinions based on Sentiment Analyses

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that deals with the interaction between human language and computers to allow them to understand incoming information and process it independently. One subdivision of NLP is called Sentiment Analysis (SA), also known as

opinion mining. Machine learning approaches to SA classify texts by identifying their sentiment based on previously learned patterns. Machine learning "addresses the question of how to build computers that improve automatically through experience" (Jordan & Mitchell, 2015, p. 255). SA is a common tool to summarize emotional communication patterns on social media and is becoming increasingly important in the field of social media analytics (Stieglitz et al., 2018). Sentiment analyses are ideally suited to address the distribution of positive and negative viewpoints on a question of interest.

This method is of particular interest for the identification and further investigation of political homogeneity online. With this approach, it is possible to recognize whether and which people expressed a positive or negative stance on an issue and whether users are referring to each other. There are only few studies on political homogeneity which use machine learning approaches to infer the political views of users from the content of their messages. Colleoni et al. (2014) classified Twitter users as either political or nonpolitical (based on training data from blog posts) and as either Democrat or Republican (based on training data from users' tweets). Their results suggest that the degree of homophily varies by political orientation: Democrats were less likely to have outbound ties to Republicans than Republicans to Democrats. Studies employing similar methodical approaches found that users are more likely interact with those who express similar views or stances than with those voicing dissimilar views (Himelboim et al., 2016; Williams, McMurray, Kurz, & Lambert, 2015).

While previous research, therefore, offers initial evidence on how expressed sentiments are spread all over a network, most previous studies investigated the Twitter network in which users are explicitly connected to each other (by the feature of "following") and this original connection might be subjected to selective exposure tendencies (i.e., getting virtually acquainted only to those who are politically similar). Still, it has been left open how users respond to each other on particular issues on platforms that have less structured networks (e.g., YouTube), increasing the chance of getting exposed to counter-attitudinal content. For this purpose, it is necessary to a) focus explicitly on discussions about specific political topics and b) analyze the network and its sub-networks that are formed based on these topical interactions.

The Present Approach: A Combination of Sentiment and Social Network Analyses to Assess Opinion-Based Homogeneity

So far, we are unaware of approaches in which homogeneity is applied to individual topics and simultaneously combined with automated content

analysis and SNAs. Previously, the determination of homogeneity was based on the basis of ideological classifications (i.e., the network patterns among liberals versus conservatives). We are only aware of few studies which examined polarization on the basis of topic-oriented approaches (e.g. Chan & Fu, 2017; Häussler, 2018). However, as public opinion forms based on issue-related discussions, it is key to focus on the analysis on specific topics. To this end, a combination of NLP – more precisely, automated SA – and SNAs is necessary. The present approach is structured as follows: First, based on manually labeled comments, the SA is performed with an SVM to predict the opinion climate for the entire network. Second, the results of the SA are then transformed into a network structure to compute the opinion-based homogeneity using the E-I index.

Method

Dataset

All data in this study were collected using a custom developed Python application which is directly connected to the YouTube API. Our application is able to collect multiple datasets by querying the internal YouTube search list, the video list, the comments list, and the replies list of each individual video. Each request to the respective list has its own URL that allows the API to be accessed and data to be collected[2]. For each list, we stored the requested data in a relational database.

The collected data contain the comments and replies of three controversial topics in Germany: "Kopftuchverbot in Deutschland" (headscarf ban in Germany), "Adoption für homosexuelle Paare" (adoption for same-sex couples) and "Klimawandel" (climate change) which also served as search queries. All of these topics are associated with political questions on which members of society have offered different answers. It has been suggested that especially morally loaded and controversial topics imply the potential to elicit processes of homogenization of opinion climates over time (Noelle-Neumann & Petersen, 2004). Accordingly, we believe that opinion-based homogeneity is more likely to be prevalent when focusing on such political topics (see Appendix A[3] for more information about these topics).

When requesting the videos via search list, the parameter "relevantLanguage" was set to the value "de" in order to get primarily German content. Furthermore, we sorted the search queries for videos according to their relevance using the parameter "order," whereas the parameter value is set to "relevance." While the two datasets "adoption rights" and "headscarf ban"

were acquired on May 15, 2018, the dataset on "climate change" was collected on January 22, 2019. Each dataset contains the user-generated comments as well as associated replies.

Table 1 provides an overview of the crawled videos with their corresponding search term and the data provided by this crawling. To analyze a more accurate selection of videos that reflect political issues, we filtered the videos by a specific categoryID[4]. In this case, a categoryID of 25 indicates the category of "Politics and News" in the YouTube API.

Table 1. Crawled YouTube videos.

Search keyword	total results	total likes	total dislikes	total views	total comments	filtered comments
Adoption for same-sex couples	266	31,876	8,509	2,576,318	15,889	8,443
Headscarf ban in Germany	320	199,912	26,393	7,247,958	48,354	14,277
Climate change	336	167,236	16,136	10,387,029	46,894	18,185

Classification of opinions in social media

Manual Labeling

We created a human-annotated gold standard to create a sample of the 4,000 German YouTube comments for each topic by defining a coding scheme. This scheme ensures that the unlabeled data can be assigned to a unique class which represents the sentiment of the message. We use the term "sentiment" referring to comments expressing a positive or negative stance towards a specific topic (e.g., if a comment states "I hate headscarves," this comment is classified as having a "negative" opinion of headscarves). This does not apply to comments whose general tone is positive or negative if they do not explicitly express a stance on the respective controversy.

We selected two well-trained independent annotators who received the same dataset with 4,000 randomly selected comments and replies for each topic. The data were labeled considering three mutually exclusive classes: negative, positive, and others. The coding scheme with corresponding topics and the listed classes is represented in Appendix A[3].

Agreement between the two annotators was measured using Krippendorff's alpha (Hayes & Krippendorff, 2007). The value of 0.63 was obtained for 3-class annotation of the adoption rights data, whereas a value of 0.67 was obtained for the headscarf ban data. In the case of the climate change dataset, a value of 0.54 was determined. All inter-annotator agreement values are valid for further processing. To ensure better results for the machine learning model, we decided to use only those comments for

further analysis on which both annotators agreed. This strategy guarantees that the sentiment can be clearly assigned to a unique class without inconsistencies. Table 2 shows the distribution of sentiment classes for each of the three datasets.

Table 2. Labeled datasets indicating the distribution of different classes.

	Dataset		
Sentiment	Adoption rights	Headscarf ban	Climate change
Negative	339	400	416
Positive	530	294	356
Others	2432	2769	2328

To derive the impact of the data showing a disagreement between the annotators (borderline cases), we later projected these data onto our trained model to determine to what degree our model takes these borderline cases into account. The contingency table and the graph can be found in Appendix B[3].

Data Pre-Processing

We implemented multiple data pre-processing steps that structure and clean the data to decrease the level of noise in the subsequent analyses. These steps were the creation of a training (80%) and a testing set (20%), their cross-validation, and the transformation of cleaned comments to Term Frequency-Inverse Document Frequency (TF-IDF) vectors (for more information about the data pre-processing see Appendix C[3]).

Support Vector Machine (*SVM*)

The application of SVM in text classification or SA has been successfully carried out in many studies. A recent study used the in-memory framework Apache Spark to apply a SA by using an SVM with an rbf kernel to classify microblog comments (Yan, Yang, Ren, Tan, & Liu, 2017). Al-Smadi, Qawasmeh, Al-Ayyoub, Jararweh, and Gupta (2018) compared the performance of recurrent neural networks (RNN) and SVMs on a comprehensive aspect-based SA of Arabic hotel ratings. The results indicate that the SVM performs superior to the deep RNN in terms of the research tasks (aspect category identification, aspect opinion target expression, and aspect sentiment polarity identification). However, the use of SVM combined with the network method to measure homogeneity/heterogeneity in online networks is novel. The results of the above-mentioned studies were very promising, and the performance of the classifiers was very high. Therefore, we decided to adapt them as a basis for our research.

The training of the SVM is realized by a pipeline (fixed sequence of steps) which starts by importing the cleaned training dataset and transforming the text data into numerical feature vectors to make them readable for the algorithm. We used a bag-of-words approach of assigning each word to an integer and returning a vocabulary dictionary in the form of a document-term matrix. The pipeline ends by fitting the TF-IDF vectors in the SVM. Combining the processes of 5-fold cross-validation and grid search, we can initialize different parameters during training and localize the best combination of parameters for each fold separately. The best parameter set is used which reaches the highest subjective F1-score.

The F1-score is used to determine the performance of the model. Especially when the class distribution is uneven, it is more precise than the simple accuracy measure. In a systematic test, we used an SVM with a linear kernel based on the LIBSVM implementation (Chang & Lin, 2011) of scikit-learn (Pedregosa et al., 2011). The optimization of the parameters was carried out through a grid search in 5-fold cross-validation. Instead of only tuning the parameters of the classifier, we also tuned parameters that deal with the process of data pre-processing. The list of all tuned hyperparameters is given in Appendix D[3].

The evaluation of the final model with their optimal parameters is based on the unseen test dataset. We apply the weighted F1-score as the metric to measure the performance of the model. Table 3 reveals the results of the prediction on the test dataset with their metrics.

Table 3. Summary of the precision, recall, F1-score for each class.

		Metrics			
Topic	Sentiment	Precision	Recall	F1-Score	Support
Adoption rights	Negative	0.62	0.52	0.56	64
	Positive	0.61	0.75	0.67	108
	Others	0.93	0.91	0.92	489
	Weighted avg.	0.85	0.84	0.85	661
Headscarf ban	Negative	0.56	0.53	0.54	76
	Positive	0.72	0.54	0.62	57
	Others	0.93	0.96	0.95	560
	Weighted avg.	0.88	0.88	0.88	693
Climate change	Negative	0.67	0.63	0.65	99
	Positive	0.49	0.41	0.44	69
	Others	0.91	0.95	0.93	452
	Weighted avg.	0.83	0.84	0.83	620

The normalized confusion matrices for the three datasets (see Figure 1) shows that the F1-score is strongly driven by the "others" category. Precision and recall for the positive and negative categories are lower. In summary, the confusion matrices show that the biggest performance losses are due to the classes positive and negative. In both classes, data is likely to be classified in the opposite category which may be due to the low amount of training data.

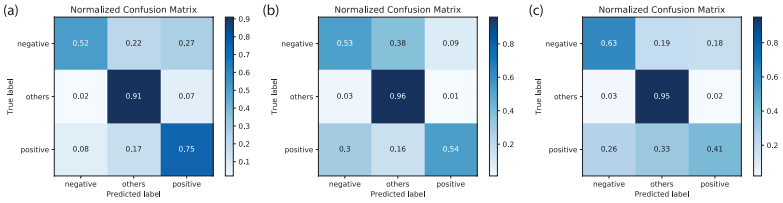

Figure 1. Normalized confusion matrix across all classes.

(a) Adoption rights

(b) Headscarf ban

(c) Climate change

Since the classifier achieves valid predictions on the test dataset and an adequate F1-score of 0.85 for the adoption dataset, 0.88 for the headscarf ban dataset, and 0.83 for the climate change dataset, it was used to predict the sentiment of the comments across the whole dataset.

Some users wrote several comments, each of which may express a combination of stances. To simplify the visualization of the network structure and the calculation of homogeneity, each user was assigned exactly one class as follows. Platt scaling was used to generate probability estimates for each class and comment (Chang & Lin, 2011). In order to summarize these values, we calculated, for each user, the average probability of each class across their comments. The user was assigned to the class that was the most likely on average. For an overview of the distribution of the predictions, please see Appendix D[3].

Building a Network on YouTube

In contrast to other social networking sites such as Facebook and Twitter, in which friendship requests and follower relationships play an integral role, the structure of the social network of YouTube users is not nearly as visible. Users can interact by commenting on videos and by commenting on other users' comments. In this study, we examined the interactions

between users' comments, associated replies, and users who uploaded the video. Thus, the focus lies on the exchange of messages between users. The SNA is structured in three parts. The first part is the creation of the network using the YouTube data to visualize the interactions across all videos (see Appendix F[3]). Statistics are used to provide a general overview of the network and to detect any conspicuous features. The second part deals with the computation of opinion-based homogeneity with the Krackhardt E-I ratio of the global network. The last part of the network analysis includes the segmentation of the networks into smaller sub-networks using the fast-greedy algorithm and the calculation of opinion-based homogeneity on a macro level (covering every comment on YouTube on that topic) as well as exchanges on the micro level (in sub-networks).

As we aim to identify the extent to which users have varying opinions on a particular topic, we decided to exclude the category "others" from the analysis as well as self-links. Topic modeling was used to gain an overview of the data that was thus discarded (Appendix E[3]). The results show that the comments in this category were off-topic and therefore do not directly contribute to the discussion between proponents and opponents on the three controversies. The removal of these off-topic posts from the network led to the creation of isolated nodes that no longer had any connections to other nodes and, therefore, had a degree equal to zero. These nodes were also deleted from the network. To ensure that the results with the class "others" are not entirely ignored, we also performed the entire analysis on all three networks including this category. The findings can be found in Appendix F.

To understand and explore the network more closely and to gain a deeper insight, we have calculated various statistics and reported the results for the three datasets in Table 4.

Table 4. Network properties.

	Datasets		
Network parameter	Adoption rights	Headscarf ban	Climate change
Nodes	536	968	626
Edges	523	1064	703
Avg. degree	0.98	1.10	1.12
Diameter	3	3	4
Max out-degree	8	18	87
Max in-degree	469	615	300
Density	0.0018	0.0011	0.0018

The characteristics of all directed networks shown in Table 4 demonstrate that the average degree is about one, suggesting that a typical user interacts with approximately one other user. In general, accounts that have uploaded a video that many other users have commented on have a higher in-degree (comments addressed to them). In addition to the in-degree, the out-degree shows which users have interacted with other users the most frequently by writing a comment. The low density values might be explained by the fact that the data originates from a real network in which the users are not linked by friendships but by their comments to each other, as well as by the high number of nodes. This pattern seems plausible in a public network where the investigation and the focus is on comments. The combination of in-degree, out-degree as well as number of nodes and edges explain the difference in the diameter.

Measuring Opinion-Based Homogeneity

One of the main goals of this study is the measurement of opinion-based homogeneity based on the sentiment of comments. To measure the degree of homogeneity, the E-I index is an appropriate choice. The formula of the global E-I Index is defined as follows:

$$EI\ Index = \frac{E - I}{E + I}$$

where E is the number of external links to a given subgroup (sentiment) and I is the number of internal links to or between nodes within that subgroup (sentiment).

The index is in a range of -1.0 to +1.0. A value of -1.0 indicates that the network is entirely homophilous with respect to the classes, i.e., all connections in the network are between members of the same class (alternatively, each connected component in the graph only involves members of the same class.). A value of +1.0 indicates an entirely heterophilous network in which there are no connections between members of the same class (i.e., a multipartite graph). In addition to measuring the global homogeneity of the network, it is possible to compute a homogeneity value for each specific class (or sentiment) to identify which sentiment has characteristics of a homogeneous interaction cluster. For example, the E-I index of the negative class would be -1.0 if all connections, both incoming and outgoing, that involve a member of the negative class were links to members of the same class. It would be +1.0 if there were no direct connections between any two members of the negative class. The index has previously been used in studies to investigate homogeneity in offline networks (Eveland &

Kleinman, 2013; Levendosky et al., 2004). To clarify the interpretation of the E-I index, Appendix F[3] shows three networks with different properties.

Identification of Communities and Extraction of Sub-Networks

The detection of sub-networks to calculate the opinion-based homogeneity of each community could give further clues about the opinion climate and possible differences between the macro and the micro level. Especially in sociology, it is necessary for many activities to identify the internal structures and groups of social networks. However, this can also be applied to online social media such as YouTube, Facebook, or Twitter in order to recognize the community structure of a network of users.

For this study, we used the fast-greedy algorithm introduced by Newman (2004) and Clauset, Newman, and Moore (2004) which is a hierarchical approach for the optimization of modularity in network analysis. This algorithm has already been applied to social network data from Twitter in several studies (e.g., Mercea & Yilmaz, 2018) and has also achieved the best results in the area of community detection based on modularity (Bello-Orgaz, Hernandez-Castro, & Camacho, 2017). The goal of this technique is to optimize the modularity to find community structures in the network. The higher the modularity score, the better is the sophisticated internal structure of the network represented. To determine the algorithm, we compared fast-greedy on a test basis with two other algorithms called Walktrap (Pons & Latapy, 2006) and Louvain (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008); the results can be found in the Appendix G[3].

Results

Figures 2-4 show the graphical representations of all three topic networks. The nodes in the network represent individual users of YouTube, i.e., users who have written comments, users who have responded to comments, and channel owners, some of whom have also written comments or replies. The color of the nodes represents their sentiment score: red for negative, green for positive, and black for channel owners who have not written any comments and are only in the dataset because they uploaded a relevant video.

Due to the aggregated probability values of the individual classes, it is easy to detect which opinion the users represent. The connections of the individual nodes to each other reflect their interaction in the form of comments. It should be noted here that this is a directed network, so it is possible to see the direction of the information flow. The hubs in the network

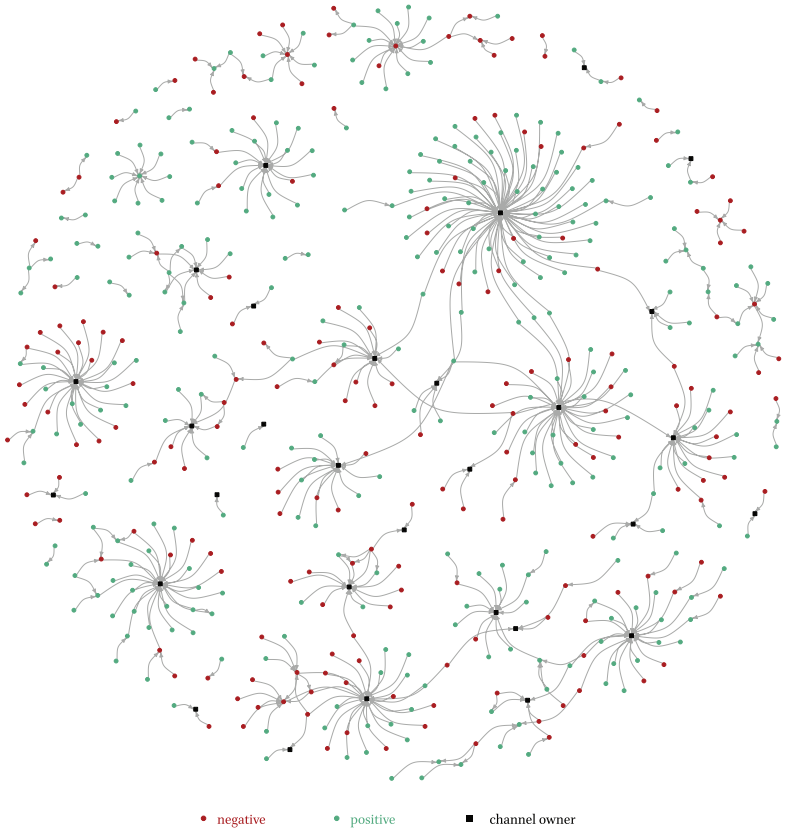

Figure 2. Discussion network on the topic of adoption rights for same-sex couples.

represent channel owners who uploaded the videos that many users commented on. Furthermore, it can be seen that apart from the hubs, the connections to the individual nodes are distributed in a very mixed way and, thus, a heterogeneous opinion climate prevails.

Looking at the classes for each topic, it is evident that YouTube users more often comment on messages that express an opinion that is different to their own than on messages with a similar stance. This is corroborated by the E-I index which approaches +1.0 and the relatively small number of internal ties (see Table 5). In addition to the visualization of the entire network, the three largest sub-networks are presented graphically in Figure 5. The visualization of the sub-networks gives a more detailed view of the network because it offers evidence about the opinion-based homogeneity related to videos with a higher number of comments.

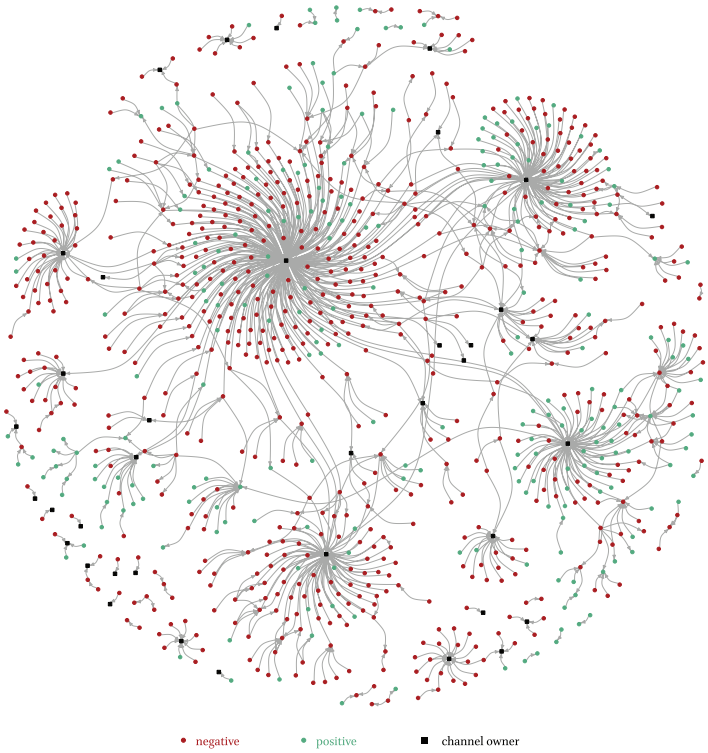

Figure 3. Discussion network on the topic of the headscarf ban.

By dividing the network into sub-networks, the individual communities can be examined more precisely, i.e., structures of single or several channel owners are recognized more effectively.

When comparing the three largest sub-networks of each dataset, it is noticeable that sub-networks on the topics "headscarf ban" and "climate change" have a higher number of users responding to comments. This is in line with the significantly higher number of comments related to those topics. Furthermore, both topics are marked by denser network structures in which different channel owners are linked by users.

The sub-communities are relatively large, and they do not reflect homogeneous opinion climates with users unanimously speaking out in favor of or against a political decision. Instead, they show a moderately diverse exchange of opinions. By examining the sub-networks, a significantly more precise analysis and results can be created for the micro-level where only

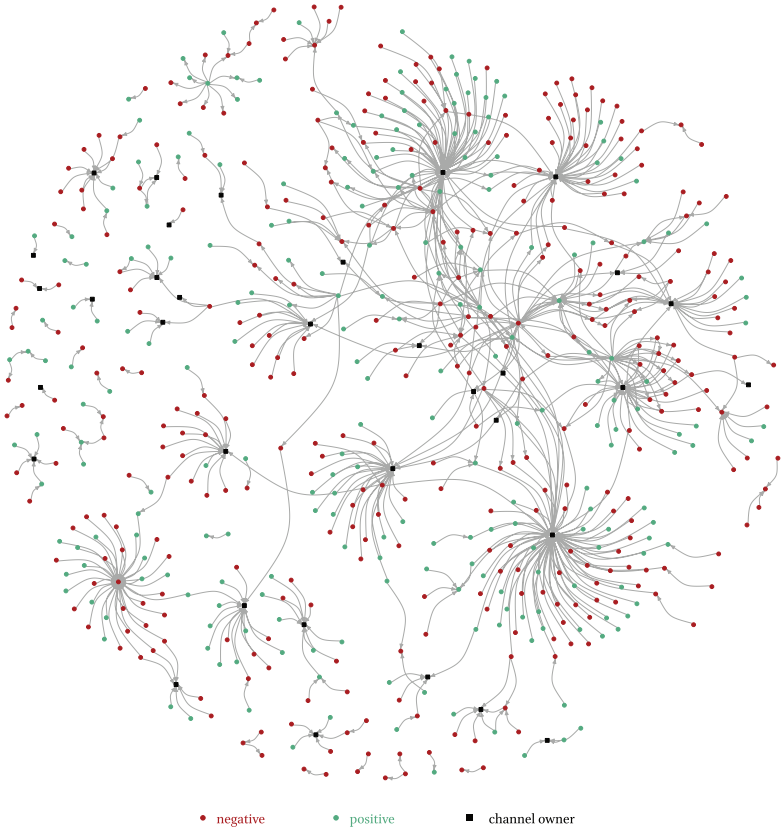

Figure 4. Discussion network on the topic of climate change.

Table 5. Properties of opinion-based homogeneity.

		Network statistics			
	Sentiment	Internal Ties	External Ties	Class E-I Index	Global E-I Index
Adoption rights	Negative	31	173	0.70	0.72
	Positive	41	278	0.74	
Headscarf ban	Negative	194	621	0.52	0.58
	Positive	28	221	0.78	
Climate change	Negative	102	320	0.52	0.61
	Positive	34	247	0.76	

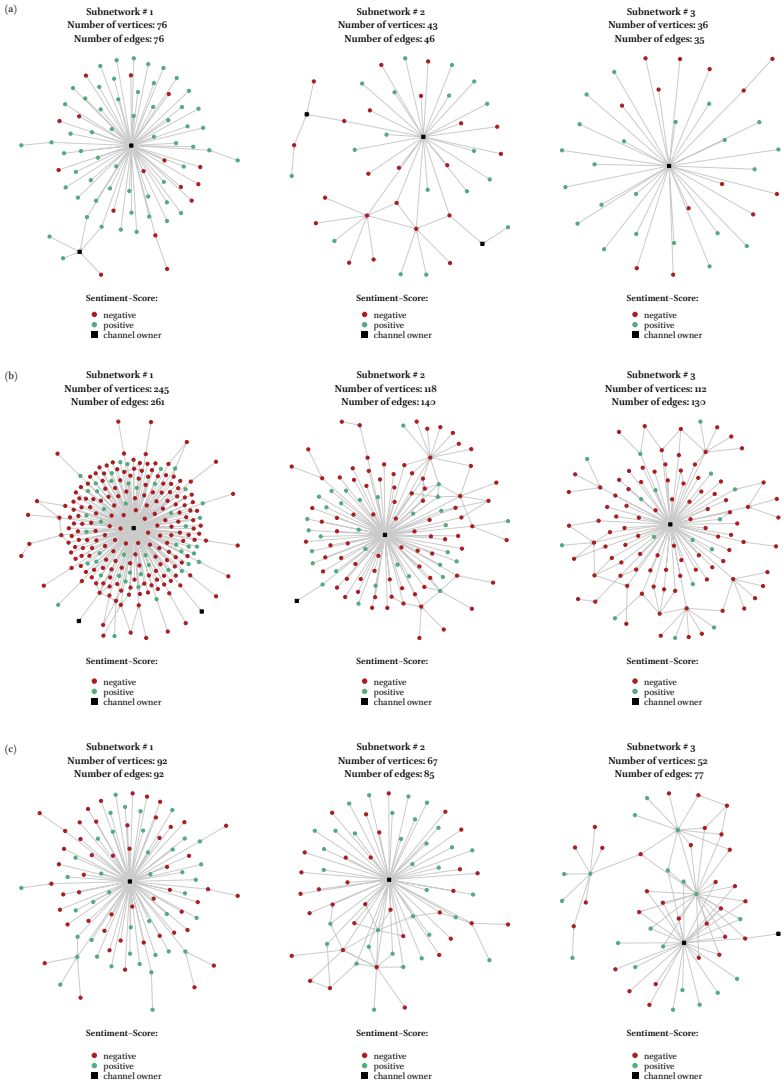

Figure 5. Sub-networks.

(a) Adoption rights

(b) Headscarf ban

(c) Climate change

specific sections of the whole network are visible. It can be noted that every sub-network exhibits heterogeneous behavior with regard to the opinion climate. Table 6 shows the results for sub-networks of global as well as class E-I Indexes.

Table 6. Properties of opinion-based homogeneity – Sub-networks.

			Statistics			
Dataset	Sub-network	Sentiment	Internal Ties	External Ties	Class E-I Index	Global E-I Index
Adoption rights	I	Negative	1	15	0.88	0.92
		Positive	2	58	0.93	
	II	Negative	9	19	0.36	0.61
		Positive	0	18	1	
	III	Negative	1	12	0.85	0.94
		Positive	0	22	1	
Headscarf ban	I	Negative	30	182	0.72	0.77
		Positive	0	49	1	
	II	Negative	28	71	0.43	0.59
		Positive	1	40	0.95	
	III	Negative	42	71	0.26	0.35
		Positive	0	17	1	
Climate change	I	Negative	2	48	0.92	0.89
		Positive	3	39	0.86	
	II	Negative	8	35	0.63	0.79
		Positive	1	41	0.95	
	III	Negative	4	26	0.73	0.61
		Positive	11	36	0.53	

Discussion

The present work was intended to (a) offer a new methodological approach to address opinion-based homogeneity using a combination of NLP and SNA and (b) provide preliminary evidence on the prevalence of opinion-based homogeneity regarding three (politically) controversial topics discussed on the platform YouTube.

Addressing RQ1, results based on the combination of NLP and SNA did not offer evidence for opinion-based homogeneity regarding positively and negatively valenced YouTube comments on the topics of adoption rights for same-sex couples, the prohibition of headscarves, or climate change. Instead, we found a moderate level of opinion-based heterogeneity when it came to the connection, that is, cross-references among user-generated comments on YouTube. In other words, comments on these three

political issues were more likely to be connected to dissimilar than to similar comments.

Regarding RQ2, it can be concluded that there are only minor differences between the macro and the micro level in the determination of opinion-based homogeneity versus heterogeneity. Both analyses – either focusing on the whole network or on sub-network – show similar structures. In particular, a closer look at the different sub-networks can lead to a more precise analysis because structures of individual communities can be focused, and opinion-based homogeneity can be calculated specifically. Given analyses at both levels, one cannot conclude that users on YouTube are exposed to a series of connected messages that all represent like-mindedness in terms of a uniform opinion climate. This result challenges previous research offering evidence for the, albeit weak, prevalence of ideological homogeneity of social networks such as Twitter (Bakshy et al., 2015; Barberá et al., 2015; Boutyline & Willer, 2017; Colleoni et al., 2014). These studies, however, focused on ideological homogeneity, that is, to what extent Democrats and Republicans interact with each other on platforms such as Twitter. Political discussions, though, may become diverse and include diverging viewpoints even within these ideological clusters. Moreover, as indicated by the same line of research, users still have ties to "the other side." While previous studies assumed that due to their cross-ideological connections, social media users might encounter content that is created or published by an ideologically deviant source (Bakshy et al., 2015; Barberá et al., 2015), it remained unclear whether users indeed encounter cross-cutting content. The present work provides initial evidence that users' opinion expressions are more likely to be associated with divergent than with congenial comments by others. In fact, this pattern is in line with the notion of "corrective action" (Rojas, 2010) stating that users feel encouraged to become outspoken online when they feel that their opinion is underrepresented. According to the patterns found on YouTube, this seems to apply as users tend to voice their political stance especially in relation to previous comments that were different to their opinion.

The only group with significantly more in-group interactions than out-group interactions, as evidenced by a negative class E-I index, is the "others" group (see Appendix F[3]). This groups consists of users that discuss topics that are only vaguely related to the controversy in question (see Appendix E[3]). From the results of the study, it appears that such comments commonly trigger a similarly off-topic response, leading to the creation of entire comment threads that diverge from the topic of the video. These groups are therefore homogeneous, but not with respect to their opinion

on the topic of the video, which would be a prerequisite for the existence of opinion-based homogeneity in the sense of the present research questions.

The combination of machine learning and SNA allowed measuring opinion-based homogeneity by assigning opinions to a particular class, training a model based on these labelled data, and applying this model to all comments. Still, it should be noted that the values predicted by means of machine learning do not reach perfect accuracy due partly to the size of the dataset and the unequal number of samples for the different sentiment classes, especially for the over-represented class "others." However, the prediction of the test datasets gives us a rough impression of the extent to which the classification works well on previously unseen data and whether the model has generalized well or only classifies examples correctly that closely resemble the training data. Looking at the performance metrics, it can be seen that the model generalizes well with class weights that are suitable for rebalanced datasets.

In general, unbalanced datasets are a common problem in machine learning contexts which can be solved by crawling and labeling even larger and more balanced datasets to improve data quality and provide more training data for the model. In the pre-labeling procedure, we have also helped to improve data quality by only using records for analysis where both annotations matched. It is remarkable that most comments crawled on all three topics did not elaborate on the question of interest. This is in line with early research evaluating the deliberative ideals of online discussions which assessed that many contributions made by users are off-topic (Janssen & Kies, 2005; Min, 2007; Schneider, 1996). Consequently, while the present findings may allow us to be optimistic about the heterogeneity of political discussions on YouTube, it raises concerns about the relative weight of these on-topic exchanges in face of a huge number of off-topic interactions.

Limitations

Our method of crawling YouTube comments about three different topics does not represent the full landscape of the political discussion on this platform but rather gives an overview of three currently discussed debates and exchanges to determine the degree of homogeneity. One reason for this is the limitation of the YouTube API which only enables crawling a fixed number of comments and videos.

Another limitation which can affect the opinion climate in the analysis, is the imbalance of the labeled classes, making the training more challenging. Using 5-fold cross validation and class weights which can be used for

addressing the generalization problem as well as for the hyperparameter search, we have tried to prevent the model from overfitting. However, one reason why the accuracy of this model is so high is that this over-represented class is more common in the training and test datasets, and it is therefore also predicted more often automatically. This also means that the accuracy of the individual models strongly depends on the available data. This, in turn, has a direct influence on the calculation of opinion-based homogeneity in the network. Increasing the amount of data would therefore also lead to the creation of a separate validation dataset which in the analysis could increase the accuracy of the model and reduce overfitting.

As a further limitation of the work, it should be mentioned that excluding the cases of disagreement between both annotators can influence the result of the classification and give a misleading impression of the accuracy of the classifier. To prevent this, a higher number of annotators would be necessary in order to have a uniform understanding of the comments and therefore increase the precision of the trained model. The results of the contingency table in Appendix B[3] show that most of these borderline cases belonged to the class "others," which is also the most frequently represented one in the dataset.

For the present study, the YouTube network was built based on the connection of videos, comments, and replies. Consequently, the network does not show the full connection structure between the individual users (e.g., friendships on Facebook). Accordingly, we consider homogeneity in the discursive sense between users although the criteria according to which the user selects individual videos or channels cannot be determined on the basis of this structure. In the present work, video uploaders assumed a key role as their opinion (provided that they expressed one) was a central connection node in the networks. Their stance was inferred from any comments they had made on their own and others' videos. Future research could also take the role of the video itself and its stance on the political question into account and investigate its interplay with the opinion climate that emerges in the related comment section.

While this applied approach has been limited to the YouTube platform, it is possible to apply the same approach to other social networking platforms such as Facebook or Twitter (using further political topics, in other language contexts) to measure opinion-based homogeneity there as well. A systematic comparison of homogeneity across different social media services will contribute to developing a robust understanding of the dynamics of political discussions online and the factors that determine whether they become homogeneous or heterogeneous.

Conclusion & Further Work

This study has developed an approach to measure opinion-based homogeneity based on textual messages with SA and SNA techniques on the YouTube platform by evaluating three relevant and politically controversial topics. Specifically, we investigated, based on communication data on YouTube, how expressed opinions in the form of user-generated comments are connected to each other and to what extent opinion-based homogeneity and heterogeneity mark the political discourse. In contrast to ideological homophily, which is more suitable for the recognition of moral values and political identities, the present approach allows the investigation of dynamic opinion climates which can change in the course of political discourses.

The combination of the two methods SNA and SA has shown that a measurement of opinion-based homogeneity based on YouTube comments is possible and can also be adapted to different topical contexts and a variety of social platforms. In the overall network, instead of finding evidence for opinion-based homogeneity, we found a moderate level of connectivity among dissimilar opinions expressed in user-generated comments. Thus, comments who expressed either a positive or a negative stance toward one of the three political issues were more likely to be associated with a heterogeneous than with a homogeneous environment. A similar pattern was found when the whole network was divided into sub-networks, e.g., in which a lot of comments were related to each other. Accordingly, this paper contributes to computational communication research in three respects:

1. It offers a blueprint for a combination of computational methods (SA and SNA) that enable the analysis of large communication datasets in light of potential social dynamics (such as communication content becoming homogeneous).
2. While previous network analyses focused predominantly on Twitter, this work relies on political communication content available on the platform YouTube, a platform that is growing as a political arena, especially for younger users.
3. Given the public debate about so-called echo chambers and political homogeneity in social media, this paper offers evidence based on automated analyses of observational data that extends previous research by not focusing on ideological homogeneity but on opinion-based and issue-related homogeneity.

Acknowledgments

This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia (Grant Number: 005-1709-0004), Junior Research Group "Digital Citizenship in Network Technologies" (Project Number: 1706dgn009).

Notes

1 https://www.similarweb.com
2 Example URL for a search query on the keyword climate change: https://www.googleapis.com/youtube/v3/search?part=snippet&relevantLanguage=de&order=relevance&maxResults=50&climate+change&key=API-KEY
3 https://osf.io/e92n3/?view_only=95ece274e9b74cc29dcadb49a06062fb
4 https://developers.google.com/youtube/v3/docs/videoCategories

References

Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., & Gupta, B. (2018). Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. *Journal of Computational Science*, *27*, 386–393. https://doi.org/10.1016/j.jocs.2017.11.006

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Beam, M. A., Hutchens, M. J., & Hmielowski, J. D. (2018). Facebook news and (de)polarization: Reinforcing spirals in the 2016 US election. *Information, Communication & Society*, *21*(7), 940–958. https://doi.org/10.1080/1369118X.2018.1444783

Bello-Orgaz, G., Hernandez-Castro, J., & Camacho, D. (2017). Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems*, *66*, 125–136. https://doi.org/10.1016/j.future.2016.06.032

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Bond, R., & Messing, S. (2015). Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook. *American Political Science Review*, *109*(1), 62–78. https://doi.org/10.1017/S0003055414000525

Boutyline, A., & Willer, R. (2017). The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology*, *38*(3), 551–569. https://doi.org/10.1111/pops.12337

Bright, J. (2018). Explaining the emergence of political fragmentation on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication*, *23*(1), 17–33. https://doi.org/10.1093/jcmc/zmx002

Chan, C., & Fu, K. (2017). The relationship between cyberbalkanization and opinion polarization: Time-series analysis on Facebook pages and opinion polls during the Hong Kong Occupy Movement and the associated debate on political reform. *Journal of Computer-Mediated Communication*, *22*(5), 266–283. https://doi.org/10.1111/jcc4.12192

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 27:1–27:27. https://doi.org/10.1145/1961189.1961199

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, *70*(6), 066111. https://doi.org/10.1103/PhysRevE.70.066111

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, *64*(2), 317–332. https://doi.org/10.1111/jcom.12084

Del Valle, M. E., & Bravo, R. B. (2018). Echo Chambers in parliamentary Twitter networks: The Catalan case. *International Journal of Communication*, *12*, 21.

Eveland, W. P., & Kleinman, S. B. (2013). Comparing general and political discussion networks within voluntary organizations using social network analysis. *Political Behavior*, *35*(1), 65–87. https://doi.org/10.1007/s11109-011-9187-4

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298–320. https://doi.org/10.1093/poq/nfw006

Geschke, D., Lorenz, J., & Holtz, P. (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), 129–149. https://doi.org/10.1111/bjso.12286

Graham, T. (2015). Everyday political talk in the internet-based public sphere. In S. Coleman & D. Freelon (Eds), *Handbook of digital politics* (pp. 247–263). Cheltenham, UK: Edward Elgar Publishing.

Guo, L., Rohde, J. A., & Wu, H. D. (2018). Who is responsible for Twitter's echo chamber problem? Evidence from 2016 US election networks. *Information, Communication & Society*. https://doi.org/10.1080/1369118X.2018.1499793

Häussler, T. (2018). Heating up the debate? Measuring fragmentation and polarisation in a German climate change hyperlink network. *Social Networks*, *54*, 303–313. https://doi.org/10.1016/j.socnet.2017.10.002

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Himelboim, I., Sweetser, K. D., Tinkham, S. F., Cameron, K., Danelo, M., & West, K. (2016). Valence-based homophily on Twitter: Network analysis of emotions and political talk in the 2012 presidential election. *New Media & Society*, *18*(7), 1382–1400. https://doi.org/10.1177/1461444814555096

Janssen, D., & Kies, R. (2005). Online forums and deliberative democracy. *Acta Politica*, *40*(3), 317–335. https://doi.org/10.1057/palgrave.ap.5500115

Jasny, L., Waggle, J., & Fisher, D. R. (2015). An empirical examination of echo chambers in US climate policy networks. *Nature Climate Change*, *5*(8), 782. https://doi.org/10.1038/nclimate2666

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kim, M. (2018). How does Facebook news use lead to actions in South Korea? The role of Facebook discussion network heterogeneity, political interest, and conflict avoidance in predicting political participation. *Telematics and Informatics*, *35*(5), 1373–1381. https://doi.org/10.1016/j.tele.2018.03.007

Knobloch-Westerwick, S. (2014). *Choice and preference in media use: Advances in selective exposure theory and research*. Routledge.

Krackhardt, D., & Stern, R. N. (1988). Informal Networks and Organizational Crises: An Experimental Simulation. *Social Psychology Quarterly*, *51*(2), 123–140. https://doi.org/10.2307/2786835

Lee, J. K., Choi, J., Kim, C., & Kim, Y. (2014). Social media, network heterogeneity, and opinion polarization. *Journal of Communication*, *64*(4), 702–722. https://doi.org/10.1111/jcom.12077

Levendosky, A. A., Bogat, G. A., Theran, S. A., Trotter, J. S., Eye, A. von, & Davidson, W. S. (2004). The social networks of women experiencing domestic violence. *American Journal of Community Psychology*, *34*(1–2), 95–109. https://doi.org/10.1023/B:AJCP.0000040149.58847.10

Lu, Y., & Lee, J. K. (2018). Stumbling upon the other side: Incidental learning of counter-attitudinal political information on Facebook. *New Media & Society*, 1461444818793421. https://doi.org/10.1177/1461444818793421

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Mercea, D., & Yilmaz, K. E. (2018). Movement social learning on Twitter: The case of the People's Assembly. *The Sociological Review*, *66*(1), 20–40. https://doi.org/10.1177/0038026117710536

Min, S.-J. (2007). Online vs. Face-to-face deliberation: Effects on civic engagement. *Journal of Computer-Mediated Communication*, *12*(4), 1369–1387. https://doi.org/10.1111/j.1083-6101.2007.00377.x

Neubaum, G., & Krämer, N. C. (2017). Monitoring the opinion of the crowd: Psychological mechanisms underlying public opinion perceptions on social media. *Media Psychology*, *20*(3), 502–531.

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, *69*(6), 066133. https://doi.org/10.1103/PhysRevE.69.066133

Noelle-Neumann, E., & Petersen, T. (2004). The spiral of silence and the social nature of man. In *Handbook of political communication research* (pp. 357–374). Routledge.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, *10*(2), 191–218. https://doi.org/10.1007/11569596_31

Prior, M. (2007). *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge, NY: Cambridge University Press.

Rojas, H. (2010). "Corrective" actions in the public sphere: How perceptions of media and media effects shape political behaviors. *International Journal of Public Opinion Research*, *22*(3), 343–363. https://doi.org/10.1093/ijpor/edq018

Schneider, S. M. (1996). Creating a democratic public sphere through political discussion: A case study of abortion conversation on the Internet. *Social Science Computer Review*, *14*(4), 373–393. https://doi.org/10.1177/089443939601400401

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168.

Sunstein, C. R. (2017). *# Republic: Divided democracy in the age of social media*. Princeton University Press.

Vaccari, C., Valeriani, A., Barberá, P., Jost, J. T., Nagler, J., & Tucker, J. A. (2016). Of echo chambers and contrarian clubs: Exposure to political disagreement among German and Italian users of Twitter. *Social Media + Society*, *2*(3), 1–24. https://doi.org/10.1177/2056305116664221

Williams, H. T., McMurray, J. R., Kurz, T., & Lambert, F. H. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, *32*, 126–138. https://doi.org/10.1016/j.gloenvcha.2015.03.006

Yan, B., Yang, Z., Ren, Y., Tan, X., & Liu, E. (2017). Microblog sentiment classification Using parallel SVM in Apache Spark. *2017 IEEE International Congress on Big Data* (*BigData Congress*), 282–288. https://doi.org/10.1109/BigDataCongress.2017.43

YouGov, & BRAVO. (2017). *Politische Jugendstudie* [*Political youth study*]. Retrieved from https://campaign.yougov.com/DE_2017_07_Political_Bravo_Jugendstudie_DE_2017_06_Political_Die_Deutschen_und_die_Politik_Landing.html

Zillmann, D., & Bryant, J. (Eds.). (1985). *Selective exposure to communication*. Hillsdale, N.J: L. Erlbaum Associates.

About the authors

Daniel Röchert, German Neubaum, Björn Ross, Florian Brachten and Stefan Stieglitz work at the University of Duisburg-Essen, Germany, Department of Computer Science and Applied Cognitive Science.

Correspondence address: University of Duisburg-Essen, Department of Information Science and Applied Cognitive Science, Group Digital Citizenship in Network Technologies; Forsthausweg 2, 47057 Duisburg (daniel.roechert@uni-due.de)

Research Paper 3: "Caught in a networked collusion? Homogeneity in conspiracy-related discussion networks on YouTube"

Type	Journal
Rights and permission	This article was published in Information Systems, 103, Röchert, D., Neubaum, G., Ross, B., & Stieglitz, S, Caught in a networked collusion? Homogeneity in conspiracy-related discussion networks on YouTube, 101866, Copyright Elsevier (2022).
Authors	**Röchert, Daniel**; Neubaum, German; Ross, Björn; Stieglitz, Stefan
Year	2022
Outlet	Information Systems (IS)
Publisher	Elsevier
Permalink/DOI	https://doi.org/10.1016/j.is.2021.101866
Full citation	Röchert, D., Neubaum, G., Ross, B., & Stieglitz, S. (2022). Caught in a networked collusion? Homogeneity in conspiracy-related discussion networks on YouTube. *Information Systems*, *103*, 101866.

Information Systems 103 (2022) 101866

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

Caught in a networked collusion? Homogeneity in conspiracy-related discussion networks on YouTube

Daniel Röchert [a,*], German Neubaum [a], Björn Ross [b], Stefan Stieglitz [a]

[a] University of Duisburg–Essen, Germany
[b] University of Edinburgh, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Article history:
Received 14 September 2020
Received in revised form 31 May 2021
Accepted 3 July 2021
Available online 3 August 2021
Recommended by Quoc Viet Hung Nguyen

Keywords:
Machine learning
Social network analysis
YouTube
Conspiracy theories
Opinion- based homogeneity

ABSTRACT

In many instances, misinformation among the population manifests itself in the form of conspiracy theories. Services such as YouTube, which allow the publication of audiovisual material in juxtaposition with peer responses (e.g., comments), function as ideal forums to disseminate such conspiracy theories and reach a massive audience. While previous research provided initial evidence about the prevalence of conspiracy theories in social media, it remains unclear how online networks discussing conspiracist content are structured. Knowledge about the network structure, however, could indicate to what extent people discussing conspiracist ideas face the risk of becoming caught in homogeneous communication cocoons. This work presents an approach combining natural language processing and network analysis to measure opinion-based homogeneity of discussion networks of three conspiracy theories (Hollow Earth, Chemtrails, and New World Order) on YouTube. A classification model was used to identify conspiracy and counter-conspiracy videos and associated user-generated comments (N = 123,642), as well as the interconnections between them. Although classification accuracy varied between the investigated conspiracy theories, our results indicated that people who expressed a favorable stance toward the conspiracy theory tended to respond to content or interact with users that shared the same opinion. In contrast, for two out of three conspiracy theories, people who advocated against the theory in their comments were more willing to engage in cross-cutting interactions. Findings are interpreted in light of the widely discussed fragmentation of homogeneous online networks.

1. Introduction

The potential threat to democratic societies of widespread misinformation in the form of conspiracy beliefs has previously been the subject of public discussions [1,2]. Research has referred to conspiracy beliefs as narratives about secret and powerful forces following plots that harm certain groups of society and benefit those forces [3]. Examples are the beliefs that the moon landing was faked by NASA, that the CIA is responsible for the John F. Kennedy assassination, or that vapor trails from airplanes (so-called Chemtrails) are sprayed by governments to manipulate the population's health [4,5]. The dissemination of conspiracy beliefs poses a hazard to individuals and societies, since exposure to material promoting conspiracy ideation decreases recipients' intentions to engage in politics [6] and pro-social activities [7]. It can also have an impact on political decisions (e.g., voting [6]) and health decisions (e.g., intention to vaccinate [8]).

Social media services such as YouTube appear to be ideal venues to circulate conspiracy beliefs among the population and insinuate a public sentiment that resonates with those conspiratorial beliefs. While there is initial evidence about the circulation of conspiracy theories in social media networks, less is known about the particular sub-networks in which conspiracist content is discussed and how these online networks are structured.

Given that conspiracy theorists (people who believe or formulate conspiracy theories) often represent minorities that face the risk of being segregated from society [9,10], it appears crucial to map the composition and interconnections of the online networks that promote such theories. Evidence about the network structure in those topical contexts could help to address the question of to what extent social media communication enables users to become caught in like-minded, that is, homogeneous clusters without exposure to cross-cutting views [11,12].

Addressing the prevalence of homogeneity in online networks, previous research indicated that in the case of networks discussing three politically relevant topics on YouTube, follow-up comments were more likely to express opposing views than similar opinions [13]. Discussions on conspiracy theories, however,

* Correspondence to: University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science, Junior Research Group "Digital Citizenship in Network Technologies", Forsthausweg 2, 47057 Duisburg, Germany.
E-mail address: daniel.roechert@uni-due.de (D. Röchert).

https://doi.org/10.1016/j.is.2021.101866
0306-4379/

might be structured differently: The spiral of silence theory [14] predicts that people are more inclined to express themselves in situations in which they feel part of the majority. Drawing on this, one could assume that supporters of conspiracy theories—as social minorities—might only voice their viewpoints in contexts in which they encounter agreement. At the same time, believing in a conspiracy theory often goes hand in hand with a need for uniqueness, that is, the wish to stand out from the mass [15], which could also lead supporters to interact with opponents of conspiracy theories. These two lines of reasoning do not really allow a prediction about the level of homogeneity within networks in which conspiracy theories are discussed: Do people only interact with each other when they agree that the conspiracy theory is valid (or true)? Addressing this question will contribute to identifying whether there are specific groups in societies that are more susceptible to becoming caught in homogeneous communication clusters filled with like-minded views.

The issue of conspiracy theories on social media has received a lot of critical attention, especially in times of the current COVID-19 pandemic, where false news is spreading on different online social media [16]. One of the greatest challenges here is correctly identifying content that contains and supports misinformation and spreads it in the form of videos and comments on social media. Using big data and machine learning methods, models can be trained to predict this content with supporting conspiracy theory content. Currently, there is no data available that links conspiracy-theory videos with conspiracy-theory comments and additionally computes their opinion-based homogeneity to be able to express conclusions about their relationships and communication pattern.

By employing natural language processing (NLP) and social network analysis, this research is intended to examine the presence of conspiracy theories and associated discussion networks on the video sharing platform YouTube. More specifically, this study analyzes: (a) The prevalence of videos that promote or debunk conspiracy theories on YouTube; (b) the social context, that is, user-generated comments, likes, and dislikes, which accompanies videos on conspiracy theories; and (c) the interconnection of discussion networks associated with those videos, that is, the opinion-based homogeneity among user-generated comments.

The paper is organized as follows: In Section 2, we explain the theoretical background of conspiracy theories in social media and their relation to the spiral of silence of homogeneous/heterogeneous groups. We present our research method consisting of the description of the dataset, the annotation of the data, the machine learning model BERT, and the network analysis in Section 3. Section 4 summarizes the results of our study and discusses them in Section 5. Finally, in Section 6, we conclude with a summary of findings and future research potential.

2. Theoretical background

2.1. Conspiracy theories in social media

The spread of misinformation in the digital age is a pressing problem that has been researched on different social media platforms such as Twitter [17,18], Facebook [2,19], and YouTube [20,21], focusing on different topics (e.g., vaccinations, rumors, and conspiracy theories). The rapid increase in users on online social networks such as YouTube or Facebook and their published content also poses risks for other users, for instance, in the form of false information, cyber bullying, and pornographic material [22]. A report by the Reuters Institute in 2019 showed that the presence and spread of misinformation are perceived—globally—as an urgent problem, especially when it comes to trusting platforms that post public content. In a survey covering 38 countries, 55% participants of all countries are concerned about not being able to distinguish between what is real and what is fake on the Internet. More specifically, 85% of Brazilians, 70% of Britons, and 67% of Americans worry about what is real and what is fake on the Internet. In Germany (38%) and the Netherlands (31%), the prevalence of concerns is lower [23]. Despite this mistrust in online platforms, the number of users following local news on the Internet is growing. According to a recent study, 89% of news is retrieved digitally, which includes news websites, apps, and social media [24].

Misinformation spread through different online channels can manifest itself in the form of conspiracy theories. While a variety of definitions of the term 'conspiracy theory' have been suggested, this work relies on the definition suggested by Keeley [25, p. 116], who referred to a conspiracy theory as a "proposed explanation of some historical event (or events) in terms of the significant causal agency of a relatively small group of persons—the conspirators—acting in secret".

With the emergence of social media channels allowing the broadcast of user-generated content and citizens' responses associated with that content, research has offered initial evidence on the role conspiracy beliefs are playing in those communication environments [26]. To analyze this content, software tools are needed; tools like HarVis provide a way to get more information on specific topics and also perform different analytics to capture, process, and visualize social media content on YouTube [27]. Focusing on the prevalence of conspiracy beliefs, Bessi and colleagues found that dealing with conspiracy pages on YouTube and Facebook goes hand in hand with users' polarization and their presence in homogeneous communication networks [1,2,28]. However, exposure to conspiracy beliefs can occur even without being polarized or deliberately seeking information on conspiracies: Allgaier [29] revealed that small variations of search terms on YouTube can lead users who are interested in science either to scientific videos or to material that promotes conspiracy beliefs framed as serious scientific evidence, making it more difficult for users to differentiate between truths and falsehoods. Another study in the context of the Zika virus outbreak in 2016 demonstrated that 12 out of 35 videos related to that topic contained conspiracy beliefs [30]. The numbers of user responses (i.e., comments, replies, likes), however, did not differ between conspiracy and non-conspiracy videos. The researchers Wood & Douglas [26] analyzed comments from several news articles on the conspiracy theory of 9/11 and found that authors of conspiratorial comments are more inclined to believe other unrelated conspiracy theories, which is in line with past research [31,32]. Furthermore, the analysis of 1459 conspiracy-supportive comments indicated a higher level of mistrust expressed than in anti-conspiracy-theory comments. Responding to the presence of conspiracy theories expressed in articles, videos, and user-generated comments in social media, there are attempts to counterargue or even explicitly debunk this kind of misinformation by creating and spreading so-called counter-messages [33–35]. An experiment showed that the rectification of misinformation in social media by presenting counter-arguments can be successful and reduce the amount of misperceptions [36]. Another study yielded mixed results, revealing that messages rejecting a conspiracy theory (e.g., on vaccination) and responding with counter-arguments can succeed, but only if these are present prior to the arguments of the conspiracy theories [37]. While counter-messages and corrective information might help to combat conspiracy theories online, Weeks & Gil de Zúñiga [35] argue that users do not commonly encounter these counter-messages "in the wild". Research on counter-messages as responses to extremist videos on YouTube, however, showed that due to the YouTube recommendation algorithm, counter-messages are directly associated with extremist videos. Therefore, those users

who view counter-videos are likely to receive a recommendation for extremist videos (which the counter-video intended to debunk in the first place) [34].

2.2. User reactions as an influential social context of conspiracy theories in social media

While the presence and spread of conspiracy theories in social media represent a societal problem per se, a comprehensive analysis needs to examine the social context in which conspiracist content is embedded [35]. As pointed out by previous research, the characteristic nature of social media is that news articles, status updates, tweets, and videos are integrated in a social context that can be made up of different types of user reactions such as likes, dislikes, or user-generated comments [38]. This line of research also argued that this social context in particular has the potential to either undo, that is, weaken, or reinforce the effects of the original content (e.g., status update or video). For instance, a YouTube video promoting the "Pizzagate" conspiracy (e.g., the debunked theory that high-ranking U.S. political officials led a global child-trafficking ring using a Pizza restaurant in Washington D.C. as their headquarters) may have a greater chance of persuading viewers when it is accompanied by a high number of likes and user-generated comments claiming that they knew that Hillary Clinton's campaign was corrupt and involved in dubious businesses.

According to the bandwagon heuristic, human beings rely on information in their environment that could reflect that a majority of, or at least many other, people agree with a certain claim and, therefore, this claim must be right (following the rule "what is popular must be good") [39,40]. Numeric information in terms of a high number of likes, thus, could serve as an important indicator for individuals, reflecting that many others approved this message, which, consequently, must be valid.

Likewise, exemplification theory [41] suggests that vivid examples of a complex issue are easier to process psychologically and, therefore, have a greater chance of affecting individuals' judgments. User-generated comments are intended to serve as these kinds of examples, concretely representing a certain stance (e.g., a pro-conspiracy theory viewpoint: "I truly believe that this child sex ring exists") or personal experience: ("I've read the Clinton emails and they clearly show this ring exists") and could be used by readers or viewers as a basis for estimates about how society might also think about this issue [42]. Following these theoretical considerations, a higher number of likes or user-generated comments associated with videos on conspiracy theories could either fortify or attenuate the persuasive effects of the original video. Given these potential effects, we are interested in examining the extent to which user reactions (number of likes, dislikes, and comments) vary between YouTube videos that promote versus challenge a conspiracy theory (Research Question 1).

However, not only the amount of user reactions might be indicative of how the social context qualifies effects of the video, but also the actual content of those reactions. While numeric information in terms of views, likes, and dislikes are perceived by users as ambiguous cues [43], a series of studies has shown that the valence of user-generated comments (e.g., supporting versus opposing a theory) as vivid exemplifications of experiences or opinions can either shape the evaluation of the original message (e.g., the YouTube video) [44], influence readers' or viewers' personal attitudes [45], or affect the opinion climate that recipients project onto the general population [42]. Against this background, we ask which opinion climate is reflected in user-generated comments (i.e., the distribution of supporting comments versus opposing comments) associated with YouTube videos on conspiracy theories (Research Question 2).

2.3. Believers in conspiracy theories: A minority in society

From a societal point of view, people who believe or express support for conspiracy theories can commonly be seen as minorities and, in many cases, as marginalized groups [9,10]. This marginalized position in society may have implications for the communication behavior of people with conspiracy beliefs. The spiral of silence theory [14] suggests that human beings are driven by the wish to be accepted by their social environment. Pursuing this goal of social approval, they feel comfortable with expressing viewpoints when they are in line with the prevailing opinion climate around them. At the same time, they withhold their personal stance when they realize that this viewpoint deviates from the mainstream, or at least from the opinion trend around them. Consequently, one could assume that individuals with conspiracy beliefs feel comfortable when discussing their views with others who also believe in the same theory and, at the same time, they avoid interactions with those who offer challenging views and could reject them for thinking differently. In the long run, this communication pattern could lead marginalized minorities to, themselves, be caught in comfortable, like-minded cocoons that solely confirm their worldview and represent a segregated cluster of homogeneous information and discussion.

Indeed, this formation of homogeneous subgroups in online communication has been a longstanding concern since the emergence of the Internet and even more salient since the rise of social media platforms [11,12]. Empirical evidence, however, has repeatedly shown that while social media users are more likely to be connected with like-minded others, they—often incidentally—interact and become exposed to content or opinionated messages that challenge their personal viewpoint or ideology [13,46–50].

The conclusion of this line of research is that especially those individuals who are politically extreme or at the margins of society are more likely to interact in homogeneous communication spaces. People that believe in conspiracy theories, however, may not always be politically extreme or members of marginalized groups [51]; therefore, it is unclear to what extent networks in which conspiracy theories are discussed are homogeneous in terms of the opinions expressed therein. Theoretically, different communication structures among conspiracy theory believers and non-believers are conceivable. Based on the spiral of silence theory [14], one could argue that those who support the validity of conspiracy theories—as a minority—would only interact with those who think alike. Research on minorities and their potential influence could challenge this view: A recent study revealed that a conspiracy mentality is often driven by the wish to stand out from the crowd and feel unique in contrast to the majority [15]. Thus, the status of being in a unique minority could be a driver to seek encounters with majority members who challenge one's conspiracy views. This, in fact, could even prove to be effective, given that minority influence research has suggested that minorities are able to modify mainstream ideas or opinions and persuade majority members by expressing their viewpoint consistently across time and situations [52,53]. If social media users with conspiracy beliefs are driven by the motive to change the opinion landscape and inject their beliefs or theories into the mainstream, it seems likely that they are going to interact with users holding and expressing diverging views in order to persuade them.

Opinion-based homogeneity is a concept that can assess the structure of discussion networks and refers to the extent to which opponents, such as believers and non-believers of conspiracy theories, might be interconnected. Specifically, this concept refers to the degree to which messages (e.g., user-generated comments) that are semantically similar are connected in the

network. Opinion-based homogeneity [13] can be measured by the global E-I index [54], which is defined as follows:

$$EI\ Index = \frac{E - I}{E + I}$$

where E is the number of external ties (ties between users and videos with different stances) and I is the number of internal ties (ties between users and videos with the same stance). The interpretation of the resulting index ranges from -1.0 to $+1.0$. In a completely heterogeneous network, a value of $+1.0$ indicates that there are no links between nodes of the same group, while in a homogeneous network, a value of -1.0 indicates that all links between nodes are connected to their specific group. A value of 0 indicates that the ties occur equally often. Additionally, to obtain a more precise picture of the homogeneity in the network, the calculation can also be performed individually for each class to compare the respective opinions in the network.

Applying the concept of opinion-based homogeneity and its operationalization to discussions of conspiracy theories on YouTube, we ask to what extent users with a core opinion for or against a conspiracy engage with the different types of videos (Research Question 3).

3. Method

In the following, we first describe the dataset that forms the basis of the analysis. We then outline two different methods for annotations: One to assess the valence of the videos and the other to identify the valence of user-generated comments. The state-of-the-art NLP deep learning model BERT (Bidirectional Encoder Representations from Transformers), was then trained with the annotated data to classify the remaining comments. Furthermore, we compared the performance of BERT with other machine learning baseline models (i.e., Logistic Regression (LR) and Support Vector Machine (SVM)), where BERT showed the best results. Using network analysis, the annotated videos and the predicted comments were linked together to build a network structure representing the discussion landscape of published videos. By measuring opinion-based homogeneity, it is possible to understand the overall role of homogeneous versus heterogeneous communication ties in the network between individual users and videos.

3.1. Dataset

To examine the presence and discussion networks of conspiracy theories on YouTube, we focused on three classic conspiracy theories that had already been investigated in previous research [55–57]. On December 22, 2018, we performed an automatic data collection on YouTube using the terms "Chemtrails Conspiracy", "Hollow Earth Conspiracy", and "New World Order Conspiracy" and filtered the query for English language content, sorted by the number of views. To assure that data collection explicitly deals with conspiracy theories, we decided to always include the word "conspiracy" in the YouTube search. Without this addition, content that was not related to the actual conspiracy theory would incorrectly be displayed. This is the case, for instance, with the search term "New World Order", where the most frequent hits are music videos by the band "New Order". We decided to focus on the videos that were viewed the most and, therefore, deliberately stopped data collection at about 100 records per topic. Furthermore, we considered the number of views, likes, or comments as an indicator of the extent to which the conspiracy theory had encountered attention on YouTube. The search term for the conspiracy theory of Hollow Earth, however, did not yield more than 89 hits. Table 1 provides an overview of the crawled videos with their corresponding search terms and the data provided by this crawling.

Table 1
Overview of YouTube data. *Note: Videos in the category "neither" are not included in the table and some videos were deleted from YouTube; therefore, the distribution of conspiracy videos is unequal.*

Conspiracy theory	Videos	Views	Likes	Dislikes	Comments	Channels
Hollow earth	59	8,630,996	65,686	10,521	24,146	51
Chemtrails	61	14,877,499	321,098	24,868	122,074	57
New world order	56	16,504,447	168,084	13,836	40,717	49

Further information on the core concepts of investigated conspiracy theories can be found in Online Appendix A.

3.2. Annotation

In this study, the annotation contains two essential components: First, the coding of the video material and its content in order to obtain information on the stance of the video, and second, the annotation of the comments and replies on these selected videos. In the following two sections, we outline the procedure in greater detail.

3.2.1. Manual video labeling

Given that it is not always possible to infer whether a video advocates in favor or against the validity of a conspiracy theory based on metadata (e.g., the title of the video), all videos were examined by three annotators according to the following classes: supporting theory, debunking theory, or neither. For the classification of the videos, a total of 75 videos per conspiracy theory were labeled (some videos have been removed from the analysis since YouTube deleted them). For the classification of the videos, the title and description of the video were considered. The minimum video length in the dataset was 37 s, while the maximum duration of a video was up to 2:29 h.

To measure the reliability of our labels, we created a smaller, likewise randomized, dataset of 40 of the 75 videos per conspiracy theory, which was then labeled by a fourth annotator. The overall percentage agreement for the Hollow Earth dataset was 50%, while the New World Order dataset reached 72.5%. The Chemtrail dataset had the highest value with an agreement of 80%. Since the reliability of the Hollow Earth dataset was comparatively low, two annotators and the first author of the paper independently went through all videos which yielded disagreement. Meanwhile, notes were documented for each video, on the basis of which the decision was justified. The preferred classes were then expressed one after the other, and, in the event of disagreement, the first author's decision was added to obtain a majority decision. Through the second round of evaluation, we were able to assign a distinct class and improve the quality of the annotation. We did not use this procedure for the other two conspiracy theories, since the intercoder agreement was satisfactory. Accordingly, we kept the classes of the first annotation for the other two datasets. The annotation of the videos is important due to the fact that it can be linked to the opinion-based homogeneity of the comments and replies to identify homogeneous spaces in the network.

3.2.2. Manual comments labeling

For the annotation of the YouTube comments, we chose the crowd-sourcing platform Amazon Mechanical Turk to annotate 8000 randomly selected comments and replies per conspiracy theory. Comments were categorized into one of three classes (pro-theory, contra-theory, other). Online Appendix B contains

Table 2
Labeled datasets with sentiment score and their numbers of samples.

Class (sentiment)	Datasets		
	Hollow earth	Chemtrails	New world order
Contra-theory	1389	2105	1215
Pro-theory	928	2383	1719
Other	5683	3512	5066
Total	8000	8000	8000

the complete coding scheme of the comments and a detailed description of the classes.

For the annotation of the data for each conspiracy theory, all annotators received rules and examples for coding the comments (see Online Appendix B). To increase the quality of the collected data, we set the Human Intelligence Task (HIT) Approval Rate (%) for all requesters' HITs greater than 95 and the number of HITs approved greater than 5000. For each comment we paid $0.01 to the annotators. To take the reliability of comment annotation into account, each comment was annotated by three annotators.

The agreement between the three annotators was measured using average pairwise percent agreement. The value of 57% was obtained for three-class annotation of the Hollow Earth data, whereas a value of 45% was obtained for the New World Order data. In the case of the Chemtrails dataset, a value of 43% was determined. To compensate for bad percent agreement, we opted for a majority vote to determine the class. To also include comments that did not yield an agreement (i.e., that were coded as a different class by all three annotators) in the analysis, we asked another well-trained annotator to label the remaining comments. This procedure ensured that all 8000 comments were used when training the model. The final distribution of the classes with their frequency is shown in Table 2.

To address the challenge of unbalanced class distribution, we included a further analysis in addition to the main analysis with the entire dataset, in which over-represented classes were undersampled to ensure an equal distribution (Online Appendix C). This not only guarantees the integrity of the subsequent results, but also ensures the comparability of the predictions of the network analyses in the later course. However, it should be mentioned that we will always refer to the entire dataset of the analyses in the further course of the work.

3.3. Bidirectional encoder representations from transformers (BERT)

In recent years, the field of NLP has changed rapidly. New deep learning approaches have significantly advanced the state of the art by achieving higher accuracy scores in various applications. We decided to use the current state-of-the-art model for NLP tasks, BERT [58], which can be used for common NLP tasks such as text classification, translation, summarization, and question-answering, and which outperformed previous machine learning techniques.

BERT, which is based on multiple transformer networks [59], uses stacked attention layers and allows training on unsupervised tasks by pre-training on a large corpus. Transformer layers allow words to be represented better in relation to all other words using self-attention to better memorize long-term dependencies in sequences. Since BERT is bidirectional and therefore uses a BiLSTM network, all parameters are represented in a way that makes them comparable to each other, allowing a higher degree of expression of the word embeddings in the corpus. In contrast to word2vec [60] and GloVe [61], which use context-free and vocabulary-based approaches, BERT represents the input as subwords of individual words that can be derived from the entire context. One of the most important advantages of BERT is its generalizability, which means that BERT models can easily be fine-tuned for various NLP tasks, especially when less data is available to solve domain-specific tasks more effectively than with conventional methods. For the training of the language model based on domain-specific data, an extra domain-specific layer is trained on the top layer of BERT using the fine-tuning process.

3.3.1. Pre-processing

We performed individual pre-processing steps on our dataset to improve the data quality and, thus, the prediction performance. Data was pre-processed differently for BERT and the baseline methods, since their requirements differ considerably. All comments and replies were converted into lowercase and hyperlinks replaced with the term "url". Furthermore, all models (BERT and baselines) were split into training (80%) and test data (20%). In general, for the baseline process, we tokenized the words. Subsequently, we converted a collection of comments to a matrix of token counts and used TF-IDF (Term Frequency–Inverse Document Frequency) to achieve a detailed word representation of important terms. This procedure was implemented in pipelines, which are fixed series of workflows of several tasks.

Using a pre-trained BERT model, the data pre-processing needs to be adapted to the model. First, we shortened the comments for all datasets to the maximum sequence length. To this end, we concentrated on the median of all comments. Since our datasets contained individual comments with a large sequence length of comments, the arithmetic mean is not appropriate. Since the median sequence lengths of the different datasets are between 85 and 130, we set the maximum length of the sequences to 128. Our trained BERT model needed more comprehensive preprocessing steps in order to be able to process the data. Therefore, we tokenized the data using the tf-hub model, which simplifies pre-processing. For this process, the words are converted to lowercase characters and then tokenized by WordPiece tokenization [62]. Therefore, words are split into small subwords, e.g., "believing" into "believe" and "###ing", which guarantees that a wider spectrum of out-of-vocabulary (OOV) words can be covered. After tokenization, the vocabulary is initialized where the most common combinations of existing words in the vocabulary are added iteratively; if words do not exist in the vocabulary, they are represented by individual characters: #H#o#l#l#o#w#E#a#r#t#h. Finally, special tokens are added at the beginning and at the end of the sentence, making it possible to find a better semantic connection between the sequences using the attention layer. For example, the token "[CLS]" marks the beginning of the sentence, while punctuation marks or the end of sentences are marked with "[SEP]".

3.3.2. Fine-tuning

For our analysis, we applied the official uncased model, which was pre-trained on Wikipedia (2.5B words) and the BookCorpus (800M words) and includes 12-layer, 768-hidden, 12-heads, and 110M parameters. This BERT model is able to predict YouTube comments and replies that contain content on conspiracy theories by classifying three classes (pro-theory, contra-theory, other). Concerning the training of the models, we set a batch size of 32, due to the fact that our dataset is not large enough and the classes are distributed unequally. Furthermore, we decided to evaluate four different epochs (1,2,3,4) in order to have comparative values within the BERT models. We set the learning rate to 2×10^{-5} with a warm-up proportion of 10% to gradually increase the small learning rate. Since this is a sequence classification task, the label probabilities are computed with a standard softmax output layer.

We applied the machine learning models Logistic Regression (LR) and Support Vector Machine (SVM) with a linear kernel

Table 3
Model evaluation of deep learning and machine learning methods on the test dataset.

Dataset	Models	Epoch	Macro			Weighted		
			Precision	Recall	F1 score	Precision	Recall	F1 score
Hollow earth	**BERT**	1	0.570	0.460	0.465	0.694	0.741	0.695
		2	**0.592**	**0.562**	**0.575**	**0.727**	**0.743**	**0.734**
		3	0.585	0.562	0.571	0.721	0.736	0.727
		4	0.591	0.537	0.554	0.718	0.743	0.726
	LR	–	0.533	0.557	0.542	0.712	0.684	0.696
	SVM	–	0.533	0.554	0.541	0.714	0.689	0.700
Chemtrails	**BERT**	1	0. 596	0.575	0.570	0.605	0.617	0.597
		2	**0.595**	**0.583**	**0.583**	**0.606**	**0.617**	**0.606**
		3	0.574	0.571	0.571	0.588	0.594	0.590
		4	0.559	0.559	0.559	0.578	0.579	0.578
	LR	–	0.552	0.549	0.549	0.568	0.575	0.570
	SVM	–	0.559	0.558	0.558	0.575	0.578	0.576
New world order	**BERT**	1	0.411	0.449	0.423	0.561	0.678	0.610
		2	0.541	0.501	0.507	0.636	0.671	0.645
		3	**0.539**	**0.507**	**0.517**	**0.634**	**0.660**	**0.643**
		4	0.531	0.508	0.514	0.633	0.656	0.641
	LR	–	0.523	0.503	0.508	0.627	0.652	0.636
	SVM	–	0.496	0.508	0.500	0.626	0.603	0.613

Table 4
Summary of the precision, recall, and F1 score for each class. Prediction based on the final BERT models (second epoch) to predict user-generated comments for each conspiracy theory.

Dataset	Sentiment	Metrics			Support	Prediction
		Precision	Recall	F1 score		
Hollow earth	Contra-theory	0.507	0.403	0.449	278	221
	Pro-theory	0.440	0.400	0.419	190	173
	Neither	0.830	0.884	0.856	1132	1206
	Weighted avg.	0.727	0.743	0.734	1600	1600
Chemtrails	Contra-theory	0.530	0.392	0.451	426	315
	Pro-theory	0.587	0.556	0.571	491	465
	Neither	0.667	0.801	0.728	683	820
	Weighted avg.	0.606	0.617	0.606	1600	1600
New world order	Contra-theory	0.361	0.249	0.295	245	169
	Pro-theory	0.514	0.434	0.470	346	292
	Neither	0.742	0.837	0.787	1009	1139
	Weighted avg.	0.634	0.660	0.643	1600	1600

based on the LIBSVM implementation [63] to a TF-IDF weighted bag of words as baseline approaches. The hyperparameter search used a grid search with five-fold cross-validation to find the best parameters. Domain-specific models were built separately for each dataset.

3.3.3. Evaluation

Table 3 shows the results from the prediction of the test dataset to compare the applied models with each other using the weighted average and macro-average metric of the F1 score. The comparison illustrates that BERT achieves the best results within the three datasets and, thus, outperforms the baseline models. A detailed illustration of the prediction within each class of BERT can be found in Table 4. In particular, it can be seen that reaching a good accuracy is more challenging for the "contra-theory" class in the New World Order dataset. However, this problem does not seem to be due to the model, since the baseline models reveal the same patterns. For this reason, one can assume that this problem is due to the unbalanced dataset and its small number of trained records and that better results could be achieved with additional datasets. Based on our analysis using the undersampled dataset, we found that the range of F1 scores between the three classes decreased. However, the results also show that the value of the weighted average F1 score decreased overall due to the data reduction (see Online Appendix C). When these results are compared to the baseline models, it is noticeable that the results of some baseline models still perform better than those of BERT models trained in only one epoch.

After the results of the BERT classifier on the test datasets were found to be good, the labels of all comments and replies to the conspiracy theories could be predicted. An overview can be found in Table 5. For the further course, we decided to use the BERT model with the two epochs for the datasets Hollow Earth and Chemtrails and with the third epoch for the New World Order dataset as a basis for further predictions. To take into account the fact that a user can write multiple comments, we considered the probabilities of each class for all written comments and calculated, for each user, the average probability of each class over their comments. Each user was assigned the class that was the most likely on average. This makes it possible to condense a user's entire communication history into a single value, which is helpful for visualization purposes. This representation of the users is especially important for building the network, as well as for the calculation of opinion-based homogeneity.

3.4. Network analysis

To calculate the opinion-based homogeneity, we converted the data into a network as follows: Each YouTube video is a node, and each user who commented on at least one of the videos is also a node. Edges represent interactions, that is, two nodes are linked by a directed edge from node A to node B if user A has

Table 5
Predicted sentiment and numbers of comments of the whole dataset with the trained BERT models.

Class (sentiment)	Dataset		
	Hollow earth	Chemtrails	New world order
Contra-theory	4900	19,437	11,999
Pro-theory	2962	17,555	13,184
Other	7450	25,239	20,916

commented on video B or if user A has replied to a comment made by user B. In the resulting network, video nodes tend to be hubs, since videos typically receive many more comments than the typical comment receives replies.

We determined the stance of each node towards the respective conspiracy theory (pro, contra, or other). The stance of video nodes had already been determined by manual annotation (see Section 3.2.1). For user nodes, the classifier outputs for their individual comments were aggregated using the arithmetic mean in order to take all their comments into account (compare [13]).

Nodes in the "other" class were removed since they were not relevant to studying the network relationships between supporters and opponents of the conspiracy theory. For the same reason, self-loops (comments on users' own videos and replies to their own comments) and isolated nodes (videos without comments and comments without replies by someone else) were removed.

We then calculated the global E-I index [54] and directed per-group E-I indices. In the calculation of per-class E-I indices, the direction of the edges was taken into account by only counting outgoing ties as external ties. As a result, the per-class index reflects the choices of the members of that group regarding who to interact with, and it therefore allows for a more accurate picture than the commonly used undirected group-wise E-I index.

To examine whether a given E-I index is significantly smaller or greater than would be expected if group members had no preference for internal or external ties, we used a permutation test (compare [64]). This sampling distribution of the E-I index is obtained by repeatedly rewiring each edge of the graph. This method keeps the number of nodes in each group constant, as well as the number of ties in the network (and thus its overall density). It thereby tests the null hypothesis that edges are distributed at random between the nodes.

To generate a network structure from the collected YouTube data and, thus, to calculate the opinion-based homogeneity, the data must first be converted. This mapping of the network allows a detailed inclusion of videos in the network that distribute a particular opinion, as well as users who respond to the video with comments. The relevant nodes, which are thus represented as hubs, are also included in the calculation of the E-I index. It is important that these hubs also have a stance, since they act as key players in the network and are likely to mark the general valence of the discussion. Users can also react to users to stimulate discussion and respond to different or similar opinions.

4. Results

The annotation of the videos on three conspiracy theories revealed that, in our YouTube dataset, the most common videos were those that supported the theory rather than debunk it. In relative numbers this means that videos on YouTube supporting conspiracy theories (58%–81%) are more prevalent than videos that oppose such theories (8%–33%). In particular, a comparison of the three conspiracy theories shows that, in the "New World Order" dataset, videos that support the theory are clearly more prevalent (81%) than those counter-arguing the theory (7.94%). The conspiracy theories "Hollow Earth" and "Chemtrails" have a similar distribution of supporting (61.54%, 58.21%), debunking (29.23%, 32.84%), and *neither* videos (9.23%, 11.17%). Fig. 1 shows the distribution of the categories in the videos, showing that in all theories, conspiracy-supportive content is more common than counter-conspiracy videos.

Table 6
Differences between conspiracy (N = 130) and counter-conspiracy (N = 46) videos. *Note: (C) represents conspiracy and (C-C) represents counter-conspiracy. *P < 0.05.*

Measured	Group	Mean	S.D.	*W*	*z*	*p*	*r*
Likes	C	3533.35	20,091.25	3093	0.61	0.73	0.046
	C-C	2076.8	5246.45				
Dislikes	C	231.25	677.07	2607.5	−0.85	0.197	−0.064
	C-C	416.57	749.47				
Comments	C	1032.36	4672.59	2617.5	−0.81	0.209	−0.061
	C-C	1146.3	1978.62				
Views	C	249,207.56	737,601.33	3639.5	−1.90	0.029*	−0.143
	C-C	165,564.33	617,125.45				

4.1. Popularity indicators of conspiracy theories

To address RQ1, altogether, N = 176 YouTube videos (130 conspiracy; 46 counter-conspiracy) are included in the analysis. Likes for the conspiracy group ranged from 0 to 224,881 (M = 3533.35, SD = 20,091.25), dislikes from 0 to 5649 (M = 231.25, SD = 677.07), views from 3 to 6,333,156 (M = 249,207.56, SD = 737,601.33), and comments from 0 to 50,705 (M = 1032.36, SD = 4672.59). For the counter-conspiracy group, the likes ranged from 0 to 34,300 (M = 2076.8, SD = 5246.45), dislikes from 0 to 3704 (M = 416.57, SD = 749.47), views from 2 to 4,192,952 (M = 617,125.45, SD = 165,564.33), and comments from 0 to 8898 (M = 1146.3, SD = 1978.62). Comparing the standard deviation in number of likes, dislikes, views, and comments shows that there exist huge differences within each group. We used an independent Mann–Whitney U test to compare popularity indicators such as likes, dislikes, comments, and views between conspiracy videos and counter-conspiracy videos. There was no significant difference in the numbers of: (a) Likes (W = 3093, z = 0.61, p = 0.73), (b) dislikes (W = 2607.5, z = −0.85, p = 0.197) and (c) comments (W = 2617.5, z = −0.81, p = 0.209), but there was a significant difference in the number of (d) views (W = 3639.5, z = −1.90, p = 0.029*). For the latter, means indicate that conspiracy videos are viewed significantly more frequently than counter-conspiracy videos. This effect, though, was small in magnitude (as specified by r). For three out of the four indicators, the significance test does not reject this null hypothesis. Therefore, although conspiracy-supportive videos are more prevalent, it seems that there is a balanced distribution of user reactions related to both conspiracy- and counter-conspiracy videos. The results of the test are shown in Table 6 (where *S.D.* is standard deviation, *W* is the Wilcoxon test statistic, *z* is the z-score, *p* is probability, and *r* is the effect size).

The distribution of the different popularity indicators (likes, dislikes, views, comments) of the three conspiracy theories is graphically summarized by the group's conspiracy and counter-conspiracy in a box-whisker plot in Fig. 2. Due to the strong fluctuations of the popularity indicators between the three different conspiracy theories, we decided to scale the data points based on the symmetric logarithm. The plot illustrates that the distribution of popularity indicators differs between the counter-conspiracy and conspiracy videos within the conspiracy theories. It shows that on average, counter-conspiracy videos on Hollow Earth and Chemtrails generate less attention, as measured by all popularity indicators, than conspiracy theory videos. Furthermore, as can be seen in the low average and median popularity

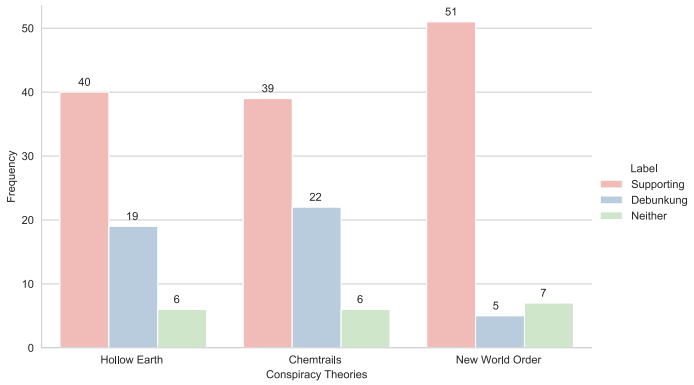

Fig. 1. Graphical representation of the distribution of conspiracy videos with their stance. Note: *Since some videos were deleted from YouTube, the distribution of conspiracy videos is unequal.*

indicators, the Hollow Earth conspiracy theory is the one that has generated the least attention. In contrast, the New World Order conspiracy theory shows that, on average, the values of the popularity indicators are higher for the counter-conspiracy videos.

4.2. Prevalence of content including conspiracy theories

To examine RQ2, we used the 8000 randomized and annotated user-generated comments to determine the distribution of the opinion climate (pro versus contra comments), which can be found in Table 2. This analysis shows that in two out of three cases, user-generated comments from supporters of the theory are more frequent than comments with a disapproving stance. However, this distribution was not found for the conspiracy theory of Hollow Earth, in which more comments included counter-messages than support of the theory.

4.3. Homogeneity and heterogeneity within discussion on conspiracy theories

Results related to opinion-based homogeneity among user-generated comments and videos on conspiracy theories (see RQ3), as the main interest of the present study, can be found in Tables 7 and 8. The results indicate that users who support the conspiracy theory are more likely to respond to videos and exchange comments with users that have the same opinion. This result can be shown by the class E-I index, which yielded negative values in all three datasets. Here, the datasets of the conspiracy theory Hollow Earth and New World Order are represented with the strongest negative E-I index values of -0.785 and -0.549, which shows relatively strong homogeneous interactions. The value of -0.221 in the dataset Chemtrails also represents a negative E-I index, but is not as strong as for the other two conspiracy theories.

Furthermore, it is noteworthy that for two out of three theories, people who advocated against the conspiracy theory show more heterogeneous communication behavior, except for the conspiracy theory Chemtrails. This is corroborated by the positive values of the E-I index: the dataset Hollow Earth has a value of 0.708, and New World Order of 0.377. The dataset of Chemtrails has a small positive value of 0.031. However, the value is close to 0 and can therefore be interpreted as neither homogeneous nor heterogeneous.

Table 7
Determining opinion-based homogeneity.

	Sentiment	Network statistics	
		Internal ties	External ties
Hollow earth	Contra-theory	168	984
	Pro-theory	673	81
Chemtrails	Contra-theory	2794	2971
	Pro-theory	6296	4015
New world order	Contra-theory	491	1085
	Pro-theory	2712	789

These findings are mainly consistent with the results of the analyses of the undersampled dataset (see Online Appendix C) and have the same tendency, indicating that the unbalanced nature of the dataset does not have a significant influence on the prediction using BERT.

Considering the permutation test, the results in Table 8 further show that the expected E-I index is also negative for the "pro-theory" class and positive for the "contra-theory" class. The difference between the observed E-I index and the expected E-I index for the class "pro-theory" is 0.999 for the Hollow Earth dataset, 0.167 for Chemtrails, and 0.178 for New World Order. For the class "contra-theory", the difference between observed and expected E-I index for the Hollow Earth dataset is 0.919, for the Chemtrails dataset 0.357, and 0.005 for the New World Order dataset.

Regarding the results of the null hypothesis test, the values of the observed E-I index are significantly closer to -1 than expected for the pro-theory class in two out of the three datasets (Hollow Earth and New World Order), and significantly closer to $+1$ than expected in only one (Chemtrails). For the contra-theory class, they are significantly closer to -1 than expected for Chemtrails, but significantly closer to $+1$ than expected for Hollow Earth.

For the graphical representation, the networks of the respective conspiracy theories are shown in Figs. 3–5, where the nodes are marked with the classes (pro-theory, contra-theory) that represent an individual user or the published video on YouTube. We used Gephi [65] and the Force Atlas 2 layout algorithm [66] to visualize the networks. Nodes with the color green represent individuals who expressed support for the conspiracy theory, while red nodes represent advocates against the theory. Furthermore, we marked the edges starting from the source nodes with their color in order to highlight the communication paths. YouTube

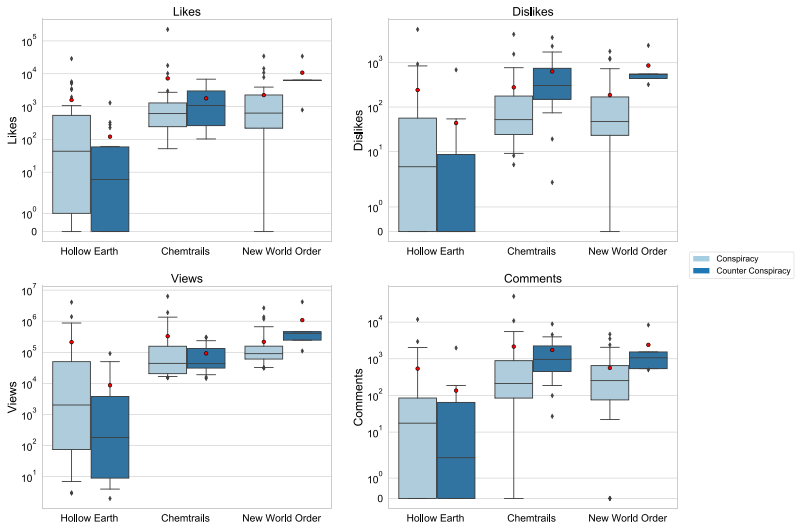

Fig. 2. Clustered box-whisker-plot of popularity indicators on counter-conspiracy and conspiracy theories. The values of the popularity indicators (likes, dislikes, views, and comments) are displayed on a logarithmic axis. The black line indicates the median, the boxes the 25th and 75th percentiles and the whiskers extend to the 5th and 95th percentiles. The categories counter-conspiracy and conspiracy are indicated with color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 8
Results of the permutation test with the observed and expected class E-I index. A permutation test with 1000 iterations was used to evaluate whether the observed index value was significantly higher ([P (obs $\geq$ exp)] or lower [P (obs $\leq$ exp)] than expected.

	Sentiment	Observed E-I index	Expected E-I index	P (obs $\geq$ exp)	P (obs $\leq$ exp)
Hollow earth	Global	0.118	−0.045	1.00	<0.01*
	Contra-theory	0.708	−0.211	1.00	<0.01*
	Pro-theory	−0.785	0.214	<0.01*	1.00
Chemtrails	Global	−0.131	−0.151	0.993	0.006*
	Contra-theory	0.031	0.388	<0.01*	1.00
	Pro-theory	−0.221	−0.388	1.00	<0.01*
New world order	Global	−0.262	−0.137	<0.01*	1.00
	Contra-theory	0.377	0.372	0.582	0.396
	Pro-theory	−0.549	−0.371	<0.01*	1.00

videos can be identified by their hub-like representation, indicating the highest in-degree. To generalize, we characterized videos that expressed support for the theory as pro-theory, while videos that disapproved of the theory were classified as contra-theory. The edges between the nodes reflect the lines of communication between the individual actors. We generated the network as a directed graph to see to whom the comments and answers are addressed. We have summarized the properties of the networks in Table 9 to provide a more comprehensive overview of the networks. The network properties show that the videos, comments, and replies related to the theory of Chemtrails make up the largest network. Similarly, the relatively high in-degree shows that there are very influential hubs in all networks. These hubs are the users who uploaded a video on a conspiracy theory and, thus, generated a lot of attention in the form of comments and replies. The out-degree indicates that the values of Hollow Earth and New World Order are very close to each other, while Chemtrails has a very high value of 412. The reason for this may be that an influential and highly active user has commented on numerous videos or has replied to several comments from other users. Furthermore, it is noticeable that the conspiracy theories

Table 9
Network properties.

Network parameter	Datasets		
	Hollow earth	Chemtrails	New world order
Nodes	1864	11,484	5000
Edges	1906	16,076	5077
Avg. degree	1.02	1.4	1.02
Diameter	2	8	3
Max. out-degree	20	412	8
Max. in-degree	374	1099	497
Density	0.00055	0.00012	0.00020

Hollow Earth and New World Order have a diameter (maximum distance between any pair of nodes) of two and three, while Chemtrails have a diameter of eight. This difference might be due to the size of the network and their different discussions. The density (i.e., the degree of interconnectedness) of our discussion networks shows a very small value in all three networks, which can be explained by the fact that the data are based on real networks and were transformed from videos and comments.

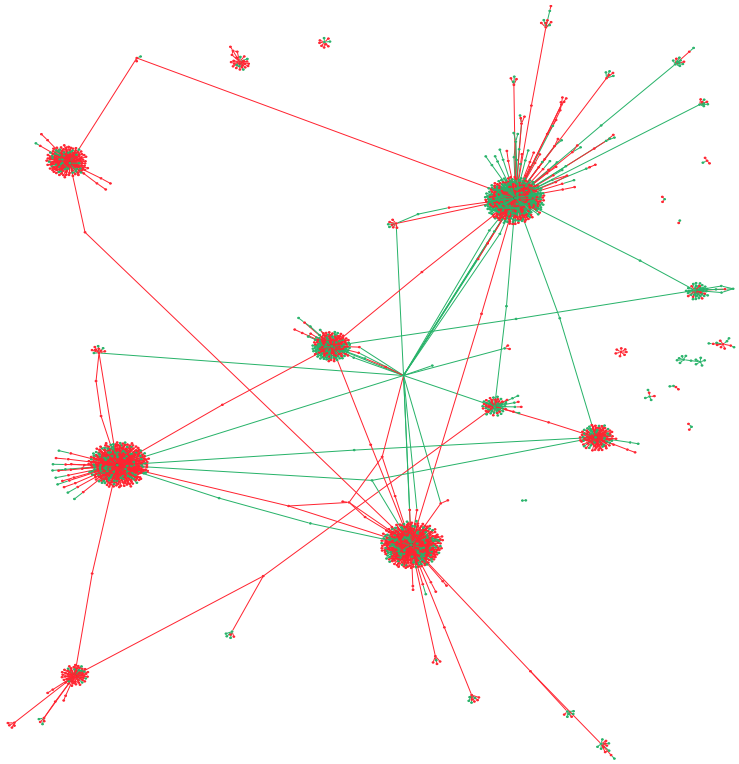

Fig. 3. Discussion network of Hollow Earth. The network has 1864 nodes, 1906 edges and an average degree of 1.02. The visualization is based on the Force Atlas 2 layout algorithm. Green nodes represent individuals who expressed support for the hollow earth conspiracy theory; red nodes represent advocates against the hollow earth theory. The edges are colored according to the color source node. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Discussion

The present work investigated the communication structure of conspiracy-related content in the form of videos and their user-generated comments on YouTube. Drawing on the spiral of silence theory and the status of a minority in a network, we were interested in the concept of opinion-based homogeneity in discussion networks related to conspiracy theories. As one of the first studies using this approach, the present work indicates that users on YouTube who express support for a conspiracy theory are more likely to engage in like-minded discussions than people who advocate against conspiracy theories.

5.1. User reactions on conspiracy and counter-conspiracy videos

With respect to RQ1, it was found that more user responses (likes, dislikes and comments) are given for conspiracy than for counter-conspiracy videos; however, these are comparatively small and statistically non-significant differences. Nevertheless, a statistically significant difference between the two groups was only found in the number of views. Previous studies evaluating conspiracy theories as misinformation observed that there is a risk that people will be indirectly influenced as a result of a high number of views, likes, or comments [67]. In this context, our findings give rise to optimism that, despite the dominance of conspiracy- over counter-conspiracy videos, the attention users pay (as measured by likes, dislikes, or comments) does not differ between the two types of videos. As mentioned in the literature review, Weeks & Gil de Zúñiga [35] suggest that countermeasures should be published by influential sources who also inspire trust and increase their social influence. Although only a small number of counter-conspiracy videos were found on YouTube, these influential videos could still contribute to the correction of conspiracy theories. Especially if these videos are characterized by a high number of popularity indicators such as likes, views, or comments, these videos could work against the narratives of conspiracy-supportive videos and reach a wide audience.

However, our results clearly showed that the number of videos featuring conspiracy-supportive content on YouTube, and thus actively contributing to the process of misinformation diffusion, is greater than the number of videos that debunked these conspiracy theories with facts. The increasing number of videos on conspiracy theories is alarming in view of the social problem of so-called filter bubbles, that is, the idea that (recommendation) algorithms shape the information landscape of users based on previous information selection patterns [34]. Thus, once users are exposed to the first video on a certain conspiracy theory, they can easily get recommendations about further videos on this theory [29]. This, in turn, could capture them in homogeneous information cocoons, reinforcing their conspiracy beliefs

Fig. 4. Discussion network of Chemtrails. The network has 11,484 nodes, 16,076 edges and an average degree of 1.4. The visualization is based on the Force Atlas 2 layout algorithm. Green nodes represent individuals who expressed support for the chemtrails conspiracy theory; red nodes represent advocates against the chemtrails theory. The edges are colored according to the color source node. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and contributing to collective polarization [68]. Therefore, given the prevalence of conspiracy videos on YouTube, more research is needed to disentangle the recommendation patterns on platforms such as YouTube in order to estimate the probability that users who were either already interested in conspiracy beliefs, were incidentally exposed to the videos, or who were actually interested in viewing theory-debunking videos will encounter even more misinformation.

5.2. Minority opinions and their homogeneous discussion network

With regard to RQ2, we found that concerning the conspiracy theories Chemtrails and New World Order, there are more comments supporting than refuting these theories. Only the dataset on Hollow Earth contains more comments against the theory than in favor of the theory. Again, this distribution provides optimistic insights into the tone of discussion on this particular topic, suggesting that promoters of this theory are met with resistance in the form of commenters who advocate against the validity of this theory. Nevertheless, it remains unclear: (a) Whether theory supporters indeed encounter and read these comments, and (b) whether counter-comments include the characteristics that are necessary to successfully outline the falsehood of this theory [36,69]. One needs to bear in mind that for two out of three cases, theory-supportive comments were more prevalent than contra-theory comments. An explanation for this could be that predominantly those who believe in the conspiracy feel the urge to discuss this theory on platforms such as YouTube. This, in turn, leads to an over-representation of support posted below conspiracy-related videos that may not represent the actual distribution of opinions among the population. Considering exemplification effects [41], this could lead to false inferences about "what most others may think" about this theory. In other words, if I see that many comments speak in favor of this theory described in a YouTube video, this could lead me to the conclusion that there is wide support in society for this conspiracy theory. This inaccurate inference could also shape my personal judgment, affecting—in the long run—my own belief in the theory [42]. Given that previous studies indicated comparatively small, albeit significant, effects of comments on public opinion perceptions, one could assume that, in the context of conspiracy theories as niche topics, effects for regular citizens are even smaller. These speculations about potential effects of encountering user-generated comments supporting conspiracy beliefs need to be addressed systematically by future research (potentially by experimental studies) to disentangle which characteristics of the comments and their readers facilitate (the perception of) public acceptance and spread of misinformation in the form of conspiracy theories. Future work may also involve using the previous data to generate simulation models that grasp the dynamics within opinion climates on different conspiracy theories.

Regarding RQ3, and consistent with the spiral of silence theory [14], users who might perceive themselves as the minority in society prefer like-minded interactions over discussions on or responses to content or comments promoting a diverging stance on the conspiracy theory. Thus, those who expressed themselves in favor of the conspiracy theories interacted in more homogeneous networks. At the same time, this finding indicates that supposed

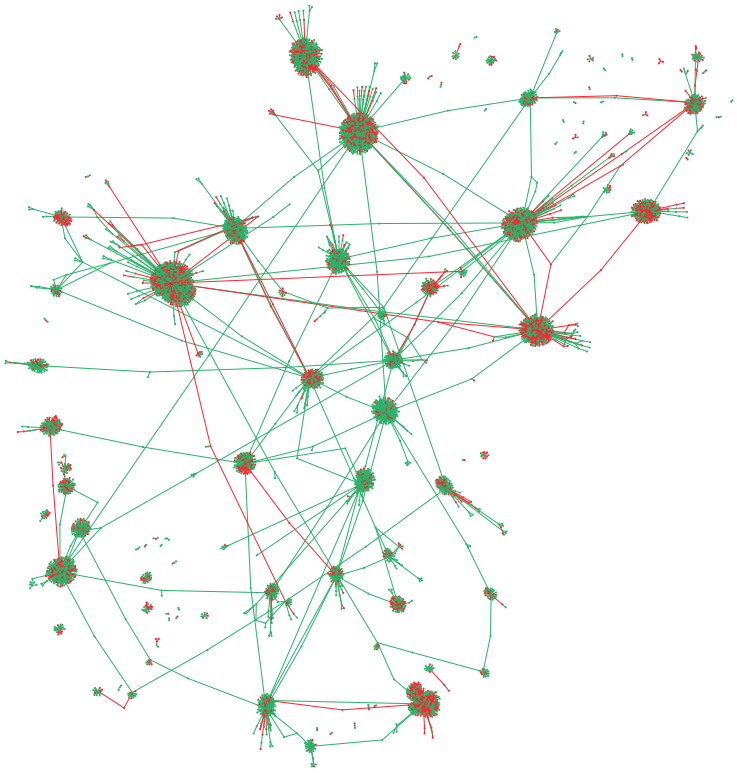

Fig. 5. Discussion network of New World Order. The network has 5000 nodes, 5077 edges and an average degree of 1.02. The visualization is based on the Force Atlas 2 layout algorithm. Green nodes represent individuals who expressed support for the new world order conspiracy theory; red nodes represent advocates against the new world order theory. The edges are colored according to the color source node. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

supporters of conspiracy theories do not exhibit behavior that lets them, for instance, contradict the mainstream in discussions [15].

In contrast, users who challenge conspiracy theories interact in more heterogeneous discussion networks covering diverse opinions and, obviously, actively seek debate. The only exception were supposed opponents of the Chemtrails theory who seem to interact in neither homogeneous nor heterogeneous opinion networks. This pattern regarding the prevalence of opinion-based homogeneity extends prior research, which largely focused on political and controversial topics [13]. For three political issues, Röchert et al. [13] found evidence for relatively heterogeneous interactions among supporters and opponents within a political debate. The discrepancy between findings might be due to the nature of the topics: Conspiracy theorists (as analyzed in the present study) are often marginalized groups [9,10] that might experience social rejection. As a special group in society, those who support these theories might feel comfortable in homogeneous, more cohesive surroundings [2]. The homogeneity among supporters of conspiracy theories has implications for the discussion about potential fragmentation of online networks [11,12]: In the long run, this potential segregation from heterogeneous interactions with challenging views could lead people with conspiracist worldviews to overestimate public support for a particular theory, feeling reinforced in their thinking. Whether this reinforcement leads to individual or collective polarization has yet to be examined by longitudinal approaches. Our findings, at least, indicate an asymmetry in the diversity of communication between those who support and those who oppose conspiracy theories. Why this pattern may vary across conspiracy theories requires further investigation by focusing on the specifics of each theory and their associated communities.

5.3. Limitations

One of the first limitations to be mentioned is the fact that only the 100 videos with the most views were crawled, which means that our data does not cover the entire discussion landscape of these topics on YouTube. In addition, the term "conspiracy" was used, which can be problematic unless people see their own theory as a conspiracy [26]. Leaving out this term would certainly have yielded more results, also covering niche networks on these conspiracies, but would also have led to many off-topic videos (e.g., "New World Order").

Due to the new YouTube guidelines in force, some videos collected in the previous step were later deleted in the course of the study. For this reason, we decided to exclude these data points from our analysis, since there are no longer any references to the original video material. Furthermore, the study shows a

static snapshot of the current YouTube landscape of conspiracy-related content with its communication network, focusing on how supporters or opponents talk about these videos. Since these are not the only three conspiracy theories on YouTube, it seems worthwhile to see whether our findings on opinion-based homogeneity can also be replicated in the context of more controversial conspiracy theories.

A further limitation of the study is the partially unbalanced dataset, which makes the prediction of some classes (pro-theory, contra-theory) more difficult than the prediction of the class "others" (neither pro nor contra), and this is also reflected in the results. It should be noted, however, that this is a representation of reality, with the majority of users writing off-topic comments. Due to the state-of-the-art language model BERT, which we used for text classification, we were able to increase the prediction accuracy and, thus, also the generalizability of our models. However, one needs to bear in mind that comments can also be predicted incorrectly. This claim is based on the primary limitation that the multiclass classification of the individual classes (especially "Contra-theory" and "Pro-theory") have a relatively low F1 score. Reasons for these low F1 scores might be, on the one hand, that not enough training data was given or, on the other, that there are similar linguistic class features. For this reason, it is not advisable to propose this particular dataset as a standard benchmark dataset. It is important to note, nonetheless, that even in the intelligent human process of labeling data, disagreements occurred. This means, firstly, that it is apparently difficult even for humans to assign comments to an unambiguous stance, and secondly, that it is even more difficult for computers to predict these manually classified comments if human coders cannot get it right.

6. Conclusion

The present study has been one of the first attempts to thoroughly measure opinion-based homogeneity in the context of three conspiracy theories on YouTube. To this end, we combined a text classification approach with BERT and a network analysis to compute the E-I index to measure the homogeneity and heterogeneity of the network based on user comments. This study showed that for three conspiracy theories users who express support for those conspiracy beliefs are more likely to interact in homogeneous networks than are users who oppose those beliefs. This pattern found within discussion networks on YouTube offers new insights for the debate on the fragmentation of social groups in online communication by specifying the (topical) circumstances under which homogeneity and potential segregation are likely to emerge. These findings have practical implications for platform providers who—by employing this methodical combination—could: (a) Detect particular sub-networks in which conspiracy beliefs are discussed and spread without any contradiction or correction, and (b) disseminate fact-checking messages in those sub-communities to counteract this ostensible legitimization of misinformation. Such an approach could help to strategically reach groups susceptible to believing in conspiracy theories in order to prevent them from becoming caught in homogeneous bubbles.

Software information

We used a Python code (Version 3.6.7) for the analysis of the YouTube comments in order to sample our datasets and train BERT models; specifically, we used the Python packages: pandas, numpy, sqlalchemy, json, requests, pickle, re, string, datetime, random, scipy, nltk, matplotlib, seaborn, itertools, scipy, sklearn, keras, tensorflow, tensorflow-hub.

For the process of network analysis and the calculation of opinion-based homogeneity, we used an R-Script (Version 3.6.2) with the following packages: igraph, tidyverse, dplyr, xlsx, compute.es, esc, pbapply.

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

The developer policy of the YouTube Data API does not permit the publication or distribution of the data used in this study. To promote the replicability of our findings, we have provided a detailed description of how to obtain the same or similar data in Online Appendix D. Please note that videos and user-generated comments used for the analyses in this study could be removed and future replications may not reach exactly the same results.

Funding

This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia (Grant Number: 005-1709-0004), Junior Research Group "Digital Citizenship in Network Technologies" (Project Number: 1706dgn009).

Appendix. Supplementary material

Supplemental material for this article is available online: https://osf.io/adu6v/?view_only=a8332f25578c428cbe333e40e634e938.

References

[1] A. Bessi, M. Coletto, G.A. Davidescu, A. Scala, G. Caldarelli, W. Quattrociocchi, Science vs conspiracy: Collective narratives in the age of misinformation, PLoS One 10 (2) (2015) e0118093.
[2] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, Walter Quattrociocchi, The spreading of misinformation online, Proc. Natl. Acad. Sci. 113 (3) (2016) 554–559, http://dx.doi.org/10.1073/pnas.1517441113.
[3] J.E. Oliver, T.J. Wood, Conspiracy theories and the paranoid style (s) of mass opinion, Amer. J. Political Sci. 58 (4) (2014) 952–966.
[4] K.M. Douglas, R.M. Sutton, D. Jolley, M.J. Wood, The social, political, environmental, and health-related consequences of conspiracy theories, Psychol. Conspir. (2015) 183–200, http://dx.doi.org/10.4324/9781315746838.
[5] V. Swami, A. Furnham, Political paranoia and conspiracy theories, power politics, and paranoia: why people are suspicious about their leaders, cambridge university press, cambridge, 2014, pp. 218–236.
[6] D. Jolley, K.M. Douglas, The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint, Br. J. Psychol. 105 (1) (2014) 35–56.
[7] S. van der Linden, The conspiracy-effect: Exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance, Personal. Ind. Differ. 87 (2015) 171–173.
[8] D. Jolley, K.M. Douglas, The effects of anti-vaccine conspiracy theories on vaccination intentions, PLoS One 9 (2) (2014) e89177.
[9] J.-W. van Prooijen, J. Staman, A.P. Krouwel, Increased conspiracy beliefs among ethnic and muslim minorities, Appl. Cogn. Psychol. 32 (5) (2018) 661–667.
[10] M.J. Wood, K.M. Douglas, Online communication as a window to conspiracist worldviews, Front. Psychol. 6 (2015) 836, http://dx.doi.org/10.3389/fpsyg.2015.00836.
[11] A. Bruns, It's not the technology, stupid: How the Echo Chamber' and Filter Bubble' metaphors have failed us, 2019.

[12] C.R. Sunstein, #Republic: Divided Democracy in the Age of Social Media, Princeton University Press, Princeton; Oxford, 2017.
[13] D. Röchert, G. Neubaum, B. Ross, F. Brachten, S. Stieglitz, Opinion-based homogeneity on YouTube : Combining sentiment and Social Network Analysis, Comput. Commun. Res. 2 (1) (2020) 81–108, http://dx.doi.org/10.5117/CCR2020.1.004.ROCH.
[14] E. Noelle-Neumann, The spiral of silence a theory of public opinion, J. Commun. 24 (2) (1974) 43–51, http://dx.doi.org/10.1111/j.1460-2466.1974.tb00367.x.
[15] R. Imhoff, P.K. Lamberty, Too special to be duped: Need for uniqueness motivates conspiracy beliefs, Eur. J. Soc. Psychol. 47 (6) (2017) 724–734.
[16] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, Antonio Scala, The COVID-19 social media infodemic, Sci. Rep. 10 (1) (2020) 16598, http://dx.doi.org/10.1038/s41598-020-73510-5.
[17] J. Shin, L. Jian, K. Driscoll, F. Bar, The diffusion of misinformation on social media: Temporal pattern, message, and source, Comput. Hum. Behav. 83 (2018) http://dx.doi.org/10.1016/j.chb.2018.02.008.
[18] M.J. Wood, Propagating and debunking conspiracy Theories on Twitter during the 2015–2016 Zika Virus Outbreak, Cyberpsychol. Behav. Soc. Netw. 21 (8) (2018) 485–490, http://dx.doi.org/10.1089/cyber.2017.0669.
[19] A. Bessi, On the statistical properties of viral misinformation in online social media, Physica A 469 (2017) 459–470, http://dx.doi.org/10.1016/j.physa.2016.11.012.
[20] Gabriele Donzelli, Giacomo Palomba, Ileana Federigi, Francesco Aquino, Lorenzo Cioni, Marco Verani, Annalaura Carducci, Pierluigi Lopalco, Misinformation on vaccination: A quantitative analysis of YouTube videos, Hum. Vac. Immunother. 14 (7) (2018) 1654–1659, http://dx.doi.org/10.1080/21645515.2018.1454572.
[21] B. Monsted, S. Lehmann, Algorithmic Detection and Analysis of Vaccine-Denialist Sentiment Clusters in Social Networks, 2019, arXiv:1905.12908.
[22] W. Kim, O.-R. Jeong, S.-W. Lee, On social Web sites, Inf. Syst. 35 (2) (2010) 215–236, http://dx.doi.org/10.1016/j.is.2009.08.003.
[23] N. Newman, R. Fletcher, A. Kalogeropoulos, R. Nielsen, Reuters Institute Digital News Report 2019, Vol. 2019, Reuters Institute for the Study of Journalism, 2019.
[24] Pew Research Center, for Local News, Americans Embrace Digital But Still Want Strong Community Connection, 2019.
[25] B.L. Keeley, Of conspiracy theories, J. Phil. 96 (3) (1999) 109–126.
[26] M. Wood, K. Douglas, What about building 7? A social psychological study of online discussion of 9/11 conspiracy theories, Front. Psychol. 4 (2013) 409, http://dx.doi.org/10.3389/fpsyg.2013.00409.
[27] U. Ahmad, A. Zahid, M. Shoaib, A. AlAmri, HarVis: An integrated social media content analysis framework for YouTube platform, Inf. Syst. 69 (2017) 25–39, http://dx.doi.org/10.1016/j.is.2016.10.004.
[28] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, Walter Quattrociocchi, Users polarization on Facebook and Youtube, PLoS One 11 (8) (2016) e0159641, http://dx.doi.org/10.1371/journal.pone.0159641.
[29] J. Allgaier, Science on YouTube: What do people find when they are searching for Climate Science and Climate Manipulation?, In 14th International Conference on Public Communication of Science and Technology (PCST), 2016.
[30] A. Nerghes, P. Kerkhof, I. Hellsten, Early public responses to the Zika-Virus on YouTube: Prevalence of and differences between conspiracy theory and informational videos, in: Proceedings of the 10th ACM Conference on Web Science, 2018, pp. 127–134.
[31] T. Goertzel, Belief in conspiracy theories, Political Psychol. (1994) 731–742.
[32] Viren Swami, Rebecca Coles, Stefan Stieger, Jakob Pietschnig, Adrian Furnham, Sherry Rehim, Martin Voracek, Conspiracist ideation in Britain and Austria: Evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories, Br. J. Psychol. 102 (3) (2011) 443–463, http://dx.doi.org/10.1111/j.2044-8295.2010.02004.x.
[33] S. Lewandowsky, K. Oberauer, G.E. Gignac, NASA faked the Moon Landing—Therefore, (Climate) science is a hoax: An anatomy of the motivated rejection of science, Psychol. Sci. 24 (5) (2013) 622–633, http://dx.doi.org/10.1177/0956797612457686.
[34] J.B. Schmitt, D. Rieger, O. Rutkowski, J. Ernst, Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube, J. Commun. 68 (4) (2018) 780–808, http://dx.doi.org/10.1093/joc/jqy029.
[35] B.E. Weeks, H. Gil de Zúñiga, What's next? Six observations for the future of political misinformation research, Amer. Behav. Sci. (2019) 0002764219878236, http://dx.doi.org/10.1177/0002764219878236.
[36] L. Bode, E.K. Vraga, In related news, that was wrong: The correction of misinformation through related stories functionality in social media, J. Commun. 65 (4) (2015) 619–638, http://dx.doi.org/10.1111/jcom.12166.
[37] D. Jolley, K.M. Douglas, Prevention is better than cure: Addressing anti-vaccine conspiracy theories, J. Appl. Soc. Psychol. 47 (8) (2017) 459–469, http://dx.doi.org/10.1111/jasp.12453.
[38] J.B. Walther, J. Jang, Communication processes in participatory websites, J. Comput.-Mediat. Commun. 18 (1) (2012) 2–15.
[39] S. Chaiken, The heuristic model of persuasion, in: Social Influence: The Ontario Symposium, Vol. 5, 1987, pp. 3–39.
[40] S.S. Sundar, A. Oeldorf-Hirsch, Q. Xu, The bandwagon effect of collaborative filtering technology, in: CHI'08 Extended Abstracts on Human Factors in Computing Systems, 2008, pp. 3453–3458.
[41] D. Zillmann, H.-B. Brosius, Exemplification in Communication: The Influence of Case Reports on the Perception of Issues, Routledge, 2012.
[42] G. Neubaum, N.C. Krämer, Monitoring the opinion of the crowd: Psychological mechanisms anderlying public opinion perceptions on social media, Media Psychol. 20 (3) (2017) 502–531, http://dx.doi.org/10.1080/15213269.2016.1211539.
[43] E.-J. Lee, Y.J. Jang, What do others' reactions to news on Internet portal sites tell us? Effects of presentation format and readers' need for cognition on reality perception, Commun. Res. 37 (6) (2010) 825–846.
[44] J.B. Walther, G. Neubaum, L. Rösner, S. Winter, N.C. Krämer, The effect of bilingual congruence on the persuasive influence of videos and comments on YouTube, J. Lang. Soc. Psychol. 37 (3) (2018) 310–329.
[45] S. Winter, Impression-motivated News Consumption–Are user comments in social media more influential than on news sites, Journal of Media Psychology 31 (4) (2019) 203–213, http://dx.doi.org/10.1027/1864-1105/a000245.
[46] E. Bakshy, S. Messing, L.A. Adamic, Exposure to ideologically diverse news and opinion on Facebook, Science 348 (6239) (2015) 1130–1132, http://dx.doi.org/10.1126/science.aaa1160.
[47] A. Boutyline, R. Willer, The social structure of political echo chambers: Variation in ideological homophily in online networks, Political Psychol. 38 (3) (2017) 551–569, http://dx.doi.org/10.1111/pops.12337.
[48] E. Colleoni, A. Rozza, A. Arvidsson, Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data, J. Commun. 64 (2) (2014) 317–332, http://dx.doi.org/10.1111/jcom.12084.
[49] E. Dubois, G. Blank, The echo chamber is overstated: the moderating effect of political interest and diverse media, Inf. Commun. Soc. 21 (5) (2018) 729–745, http://dx.doi.org/10.1080/1369118X.2018.1428656.
[50] C. Vaccari, A. Valeriani, P. Barberá, J.T. Jost, J. Nagler, J.A. Tucker, Of echo chambers and contrarian clubs: Exposure to political disagreement among German and Italian users of Twitter, Soc. Media+ Soc. 2 (3) (2016) http://dx.doi.org/10.1177/2056305116664221.
[51] B.R. Warner, R. Neville-Shepard, Echoes of a conspiracy: Birthers, truthers, and the cultivation of extremism, Commun. Q. 62 (1) (2014) 1–17.
[52] S. Moscovici, E. Lage, M. Naffrechoux, Influence of a Consistent Minority on the Responses of a majority in a color perception task, Sociometry 32 (4) (1969) 365–380.
[53] W. Wood, S. Landgren, J.A. Ouellette, S. Busceme, T. Blackstone, Minority influence: a meta-analytic review of social influence processes, Psychol. Bull. 115 (3) (1994) 323–345, http://dx.doi.org/10.1037/0033-2909.115.3.323.
[54] D. Krackhardt, R.N. Stern, Informal networks and organizational crises: An experimental simulation, Soc. Psychol. Q. 51 (2) (1988) 123–140, http://dx.doi.org/10.2307/2786835.
[55] A. Spark, Conjuring order: the new world order and conspiracy theories of globalization, Sociol. Rev. 48 (2_suppl) (2000) 46–62.
[56] D. Stupple, A. Dashti, Flying saucers and multiple realities: A case study in phenomenological theory, J. Popul. Cult. 11 (2) (1977) 479–493, http://dx.doi.org/10.1111/j.0022-3840.1977.00479.x.
[57] A.F. Wilson, The bitter end: apocalypse and conspiracy in white nationalist responses to the Islamic State attacks in Paris, Patterns Prejud. 51 (5) (2017) 412–431, http://dx.doi.org/10.1080/0031322X.2017.1398963.
[58] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Aidan Jones, Łukasz Kaiser, Illia Polosukhin, *u.a.*, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008, arXiv:1706.03762.
[60] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed Representations of Words and Phrases and their Compositionality, 2013.
[61] J. Pennington, R. Socher, C.D. Manning, GloVe: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[62] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, Jeffrey Dean, Google's multilingual neural machine translation system: Enabling Zero-Shot translation, Trans. Assoc. Comput. Linguist. 5 (2017) 339–351, http://dx.doi.org/10.1162/tacl_a_00065.
[63] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27, http://dx.doi.org/10.1145/1961189.1961199.

[64] J. Scott, P.J. Carrington, The SAGE Handbook of Social Network Analysis, SAGE, London, 2011.
[65] M. Bastian, S. Heymann, M. Jacomy, et al., Gephi: an open source software for exploring and manipulating networks., ICWSM 3 (1) (2009) 361–362, Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/13937.
[66] M. Jacomy, T. Venturini, S. Heymann, M. Bastian, Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software, PLoS One 9 (6) (2014) e98679.
[67] K.M. Douglas, R.M. Sutton, The hidden impact of conspiracy theories: Perceived and actual influence of theories surroanding the death of princess diana, J. Soc. Psychol. 148 (2) (2008) 210–222, http://dx.doi.org/10.3200/SOCP.148.2.210-222.
[68] C.R. Sunstein, Is social media good or bad for democracy?, SUR-Int. J. Hum Rights. 27 (2018) 83.
[69] G. Orosz, P. Krekó, B. Paskuj, I. Tóth-Király, B. Bőthe, C. Roland-Lévy, Changing conspiracy beliefs through rationality and ridiculing, Front. Psychol. 7 (1525) (2016) http://dx.doi.org/10.3389/fpsyg.2016.01525.

Research Paper 4: "The Networked Context of COVID-19 Misinformation: Informational Homogeneity on YouTube at the Beginning of the Pandemic"

Type	Journal
Rights and permission	This article was published in Online Social Networks and Media, 26, Röchert, D., Shahi, G. K., Neubaum, G., Ross, B., & Stieglitz, S., The Networked Context of COVID-19 Misinformation: Informational Homogeneity on YouTube at the Beginning of the Pandemic, 100164, Copyright Elsevier (2021).
Authors	**Röchert, Daniel**; Shahi, Gautam Kishore; Neubaum, German; Ross, Björn; Stieglitz, Stefan
Year	2021
Outlet	Online Social Networks and Media
Publisher	Elsevier
Permalink/DOI	https://doi.org/10.1016/j.osnem.2021.100164
Full citation	Röchert, D., Shahi, G. K., Neubaum, G., Ross, B., & Stieglitz, S. (2021). The Networked Context of COVID-19 Misinformation: Informational Homogeneity on YouTube at the Beginning of the Pandemic. *Online Social Networks and Media*, *26*, 100164.

Online Social Networks and Media 26 (2021) 100164

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.journals.elsevier.com/online-social-networks-and-media

The Networked Context of COVID-19 Misinformation: Informational Homogeneity on YouTube at the Beginning of the Pandemic

Daniel Röchert [1,*], Gautam Kishore Shahi [1], German Neubaum [1], Björn Ross [2], Stefan Stieglitz [1]

[1] University of Duisburg-Essen, Duisburg, Germany
[2] The University of Edinburgh, Edinburgh, United Kingdom

ARTICLE INFO

Keywords:
COVID-19
Misinformation
Network Analysis
Deep Learning, Social Media
YouTube
Homogeneity
Infodemic

ABSTRACT

During the coronavirus disease 2019 (COVID-19) pandemic, the video-sharing platform YouTube has been serving as an essential instrument to widely distribute news related to the global public health crisis and to allow users to discuss the news with each other in the comment sections. Along with these enhanced opportunities of technology-based communication, there is an overabundance of information and, in many cases, misinformation about current events. In times of a pandemic, the spread of misinformation can have direct detrimental effects, potentially influencing citizens' behavioral decisions (e.g., to not socially distance) and putting collective health at risk. Misinformation could be especially harmful if it is distributed in isolated news cocoons that homogeneously provide misinformation in the absence of corrections or mere accurate information. The present study analyzes data gathered at the beginning of the pandemic (January–March 2020) and focuses on the network structure of YouTube videos and their comments to understand the level of informational homogeneity associated with misinformation on COVID-19 and its evolution over time. This study combined machine learning and network analytic approaches. Results indicate that nodes (either individual users or channels) that spread misinformation were usually integrated in heterogeneous discussion networks, predominantly involving content other than misinformation. This pattern remained stable over time. Findings are discussed in light of the COVID-19 "infodemic" and the fragmentation of information networks.

1. Introduction

Social media such as Facebook, Twitter, and YouTube play a paramount role in today's society for exchanging information, especially in times of a global pandemic that forces many to stay at home [1]. This information includes latest status reports on the disease and thus helps citizens to make informed decisions about their actions in daily life. In addition to these day-to-day communications, social media platforms also provide effective channels for authorities to disseminate risk messages [2] and for members of the public to ask for help [3]. However, the new and multiple communication channels offered by social media also allow misinformation to flourish [4], which poses a potential threat to our collective health and democracy [5,6]. According to a recent poll by the Pew Research Center, 30% of U.S. adults who were primarily seeking information through social media have received "a lot" of conspiracy theory news alleging that the pandemic was deliberately planned [7].

Ever since the beginning of the pandemic, there has been a flood of myths and false reports about the virus (e.g., eating garlic prevents infection with COVID-19[1], and COVID-19 spreads via 5G mobile networks)[2]. The World Health Organization (WHO) speaks of an "infodemic" and has warned of the threat of "an overabundance of information—some accurate and some not—that makes it hard for people to find trustworthy sources and reliable guidance when they need it" [8, p. 2]. Since content on networking platforms such as YouTube is in the public domain, it is particularly important that the medical information provided and widely consumed by citizens is accurate and of high quality [9]. This can sometimes be a challenge since scientific findings related to such a complex and multi-layered issue like a global pandemic are elusive and, given the accumulation of scientific knowledge at an accelerated pace, fast-changing [10]. Thus, the dynamic

* Corresponding Author. Daniel Röchert, University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science, Junior Research Group "Digital Citizenship in Network Technologies"; Forsthausweg 2, 47057 Duisburg
E-mail address: daniel.roechert@uni-due.de (D. Röchert).

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters#garlic
[2] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters#5g

https://doi.org/10.1016/j.osnem.2021.100164
Received 25 November 2020; Received in revised form 11 March 2021; Accepted 8 August 2021
Available online 30 August 2021

nature of scientific knowledge and its recurrent effects on policymakers and citizens offers a breeding ground for the formation and spread of misinformation [6].

In relation to global public health emergencies such as the outbreak of epidemics and pandemics, previous studies addressed the presence of misleading content on the outbreaks of Ebola [11], Zika [12], and H1N1 [13]. A number of published studies have recently already addressed the spread of misinformation about the COVID-19 pandemic on Twitter [14–17], Instagram [18], and YouTube [19]. Since misinformation about COVID-19 appears to be a phenomenon across different social media platforms, the risk of users being exposed to such information appears to be continuously prevalent. Clearly, the effects of misinformation can be harmful: When reading or viewing falsehoods, for instance, about the origin of COVID-19 or the ultimate effectiveness of masks, individuals may decide to not protect themselves or others, ignoring recommendations by centers for disease control and contributing to the further spread of the infectious disease [20]. The impact of misinformation in relation to public health crises can become even more amplified when it is spread in homogeneous clusters in which false information is treated as "normal" and accurate information is absent [21]. In an era in which information and communication networks are assumed to be fragmented (i.e., divided into different groups) along ideological lines [22,23], it is conceivable that social media technologies unite individuals who believe in misinformation and, therefore, interact mainly within like-minded cocoons where information that contradicts falsehoods does not receive any attention. In light of the potential clustering of information networks within widely used platforms such as YouTube, it seems crucial to not only assess the prevalence of misinformation, but also to analyze the network in which misinformation is disseminated and discussed.

Drawing on the notion of fragmented information networks in social media, the present study introduces the concept of *informational homogeneity* to refer to the extent to which misinformation (vs. non-misinformation) is directly connected to other pieces of misinformation in a network. By relying on this concept and focusing on the increasingly popular video-sharing platform YouTube as a news source, the present study is intended to: (a) provide knowledge about the presence of misinformation related to COVID-19 on YouTube, (b) estimate the extent to which pieces of misinformation are connected among each other, and (c) analyze to what extent informational homogeneity as an indicator for fragmentation varies over time. To this end, this study analyzes a dataset of 2,585,367 comments and 10,724 videos related to COVID-19 gathered on YouTube in the period between January and March 2020 (representing the beginning of the pandemic). The analysis combines methods from deep learning and social network analysis, allowing insights into how different types of information are connected with each other in communication networks.

The paper is organized as follows: In Section 2, we explain the theoretical background of misinformation on social media in the age of the coronavirus pandemic and its relation to the fragmentation of informationally homogeneous/heterogeneous groups. We present our research approach, consisting of data collection, annotation of the data, the deep learning language model BERT, error analysis, and network analysis in Section 3. Section 4 summarizes the study results, while these are discussed in Section 5. Finally, in Section 6, we conclude with a summary of findings and future work.

2. Theoretical background

2.1. . Misinformation on social media

The pandemic outbreak of the severe acute respiratory syndrome coronavirus (SARS-CoV-2) that causes the disease COVID-19 has permanently changed the lives of millions and thus, society, in various respects. As of September 6, 2021, approximately 220 million cases of COVID-19 and 4.6 million deaths had been reported, thereby posing an enormous challenge for countries and their healthcare systems in their fight against the spread of the virus [24].

Social media have an essential function in the distribution of news during crises, since they are capable of reaching a large number of people in a short time [25,26]. In particular the information communicated by health authorities on the current status of the virus and its spread in the respective countries is an important component of prevention measures. According to previous studies, communication via social media can help to inform the public with risk messages, optimize decision-making processes [2], and ensure rapid dissemination of scientific information [27].

Making sense of the news in extreme events is a collective process; however, establishing a common consensus could also have serious consequences, especially if users are only indirectly involved in the events. If they are not well informed, this could cause rumors to arise and spread [28]. With regard to events such as the COVID-19 pandemic, Mirbabaie et al. [29] found that in particular "information-rich actors" (e.g., media organizations, emergency management agencies) are influential in social networks and that they therefore play a key role in reducing mistrust. The quest to disseminate fast-changing scientific knowledge about an urgent matter such as a global pandemic is directly linked to dealing with the emergence of misinformation, falsehoods, rumors, and misleading content [10,21].

Even at the beginning of the pandemic, the WHO recognized another problem besides the spread of the virus, i.e., the massive amount of information that could not be guaranteed to come from trustworthy and reliable sources, and defined this as an "infodemic" [8]. According to a survey, 48% of adult US Americans had already been exposed to misinformation about COVID-19 by mid-March 2020 [30].

In general, the content of political misinformation on social platforms represents a potential threat both to democratic systems and to global health. With regard to its effects on democracy [31,32], studies showed that misinformation about current events spreads faster and more widely than true information [17,33], which could lead to political misperceptions (i.e., false or inaccurate beliefs about politics [34]). In fact, the identification of misinformation is a challenge since messages mutate and are duplicated in different contexts as time goes by [35]. Misinformation related to global health issues, for instance, in the form of conspiracy theories about vaccines, has serious consequences such as reducing people's vaccination intentions and increasing distrust on this issue [36]. To counteract this, evidence-based corrections employed by algorithms can serve as preventive measures [37]. However, when misinformation is deeply rooted in people's beliefs, it is difficult to counteract [38], especially if this misinformation is embedded in communities that deal exclusively with misinformation and are more self-contained [39].

Initial studies have already examined the emergence of misinformation during the COVID-19 pandemic. Fact-checking websites have analyzed the misinformation across multiple social media platforms, most notably YouTube, Twitter, Facebook, Instagram, etc., with the rise of the pandemic over time, and the misinformation also increases at the same rate across the world in multiple languages [40,41]. Further studies also report the rise of misinformation during the beginning of the pandemic and lockdown across numerous countries, followed by a sudden decrease in misinformation. After investigating misinformation on Facebook, Twitter, and YouTube regarding the current COVID-19 pandemic, Brennen et al. [42] were able to illustrate that while the greatest share of misinformation is disseminated by ordinary people in the social sphere, this share also seems to attract the least engagement. Kouzy et al. [16] analyzed a sample of tweets based on eleven COVID-19-related hashtags and three key terms ("Corona," "Coronavirus," and "COVID-19") on February 27, 2020, and found that Twitter accounts with a low number of followers or an unverified status were more likely to spread misinformation than verified accounts and those that had more followers. Recent studies also indicate that the dissemination of misinformation seems to be platform-dependent and that the

spread of misinformation is related to the respective users of those platforms [43]. They found that the highest level of interaction between comments and posts was on YouTube and Twitter, while the distribution of user activities (reaction dynamics and content consumption) was a commonality that was similar across all platforms. Another study examined a snapshot of the most-watched YouTube videos (N = 69) on COVID-19 and found that more than a quarter of these videos contained misleading information [19]. However, based on the small size of the sample and the limited time period it covered, it is difficult to generalize the prevalence of misinformation to all of the content that is available on YouTube. Initial evidence showed that videos on COVID-19 that contained misinformation were associated with a significantly higher number of comments that also featured misinformation [44]. The service YouTube has recognized the ongoing presence of misinformation and intends to remove content that does not adhere to its guidelines[3]. Nevertheless, due to its potential global health consequences, it seems urgent to investigate the prevalence of misinformation on YouTube related to a health issue such as COVID-19—not only on one specific day but based on a longer period of time. To comprehensively assess the presence of misinformation, it is important to not only analyze the videos but also the associated comments sections:

RQ1: What is the proportion of videos and comments that spread misinformation on YouTube in the context of the COVID-19 pandemic?

2.2. . The (informational) homogeneity in online networks

Since misinformation has become a pressing issue in the agenda of social media research [45,46], scholars proposed also taking into account the networked context in which misinformation is embedded [21,47]. These proposals address the notion that misinformation could have detrimental effects on individual actions and group dynamics if it spreads in homogeneous networks in which the misperception that the misinformation is accurate is reinforced and validated by many like-minded voices in the absence of any contradiction or correction. The juxtaposition of mass media content (e.g., news coverage) and interpersonal communication (e.g., exchanges in user-generated comments) in social media could lead to even accurate (health) information promoted by news coverage being misinterpreted or mistrusted by what readers/viewers read in the comments section [48]. Therefore, analyses of the informational homogeneity in online networks need to take into account both the main media content (e.g., journalistic videos) and corresponding comment threads.

In social media, users can choose their information sources and interaction partners in a self-determined way; selective and biased information gathering is possible because people share information without verifying it [49]. Drawing on the idea of homophily as "the principle that a contact between similar people occurs at a higher rate than among dissimilar people" [50], we propose the concept of *informational homogeneity*, which refers to the extent to which uniform types of information are connected to each other. In the context of misinformation, informational homogeneity would be high if actors who spread misinformation are closely connected to each other (forming an information cluster), while they are largely disconnected from non-misinformation (which could potentially contradict or correct the misinformation).

The level of homogeneity within online networks has already been addressed by a body of research focusing on ideologies or political opinions: While a series of studies showed that people are more likely to be connected to those who are ideologically alike [51–53], a more nuanced approach focusing on homogeneity at the topic level revealed that discussion networks are more heterogeneously structured than assumed by public concerns [54]. More specifically, on YouTube, dissimilar expressions of opinion in the form of user-generated comments were more likely to be connected to each other than comments that were similar in their stance towards a topic.

The level of homogeneity within networks is not only applicable to political views but also to the accuracy of information. Following this logic, it seems conceivable that pieces of inaccurate information are directly associated with further pieces of false information. There is reason to assume that this is prevalent in social media platforms. Recommendation algorithms, like those present on platforms such as YouTube, could lead users who initially followed a video recommendation with false or inaccurate information to further content that promotes misinformation, thereby catching those users in an information network (a "rabbit hole" [55]) predominantly comprising misinformation. Indeed, a study on YouTube found that users' individual search history is responsible for recommending them misinformation content [56]. Furthermore, it was found that videos about vaccinations that contained misinformation are promoted and thus lead the user to more misinformation. On Twitter, the findings of Shin et al. [57] suggest that the dynamic communication of political rumors (misinformation) spreads in virtual cocoons. More specifically, their network analysis revealed that polarized communities of users with the same political orientation have formed and selectively spread rumors about opposing candidates. Consequently, recommendation algorithms based on users' previous interests could even amplify the effects of misinformation by conveying users the impression that there is a whole legitimate network that promotes and discusses this kind of (mis)information [55].

Despite this initial evidence on the context of misinformation in social media, it remains unclear to what extent different pieces of misinformation are linked to each other in online networks: A recent analysis of YouTube content (videos and comments) featuring misinformation in the form of conspiracy theories suggested that there is a moderate level of opinion-based homogeneity among those nodes in the YouTube network that express a stance in support of the respective conspiracy theory [80]. While this evidence on conspiracy theories may suggest that misinformation is moderately connected to further misinformation in online networks, it is unclear whether this also applies to issues relying on fast-changing evidence such as the COVID-19 pandemic. It seems conceivable that the global uncertainty related to this pandemic has led to a stronger spread of misinformation, which also diffuses into networks with predominantly accurate information. Therefore, we ask:

RQ2: How high is the prevalence of informational homogeneity of misinformation in the context of the COVID-19 pandemic?

2.3. . The fragmentation of information networks over time

The idea that news or information could spread among certain groups of people but not among others has been best described by the term "fragmentation" of news media [58]. This has been associated with the risk that communication landscapes are segmented and divided into sub-groups that are homogeneous in terms of what kind of information they receive and discuss, but also disconnected from the other sub-groups, leading to an asymmetrical diffusion of news and information [59]. From a normative point of view, fragmentation of news channels on social media, on the one hand, can have a positive effect on the distribution of relevant information since more sources of information are available [60]. On the other hand, fragmentation also carries risks and dangers, especially when these fragmented groups polarize and spread extreme ideologies, misinformation, or hate speech [23].

In direct association with the concept of homophily, studies have examined to what extent a divergence of political ideologies is responsible for fewer interactions among individuals, resulting in a fragmentation of information and discussion networks [22,61]. Empirical evidence, however, showed that the actual division in communication only applies to the politically extreme—there are still cross-cutting interactions among those who have different political views [22]. Likewise, an analysis of audience segments across different media outlets

[3] https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/

revealed a significant overlap of media consumers between all of these channels, refuting the idea of enclaves in communication networks [62]. With the diffusion of algorithms in people's communication practices, the idea of news audience segmentation has gained renewed relevance [63]: Indeed, a bounded confidence model revealed that algorithm bias in the flow of information can strengthen the fragmentation of information consumers and their opinion polarization [64].

In the context of misinformation, the fragmentation of subgroups marked by informational homogeneity would mean that certain segments of a network are disproportionately exposed to misinformation, while at the same time being disconnected from sources of accurate information. Such a network structure could lead those groups that are homogeneously exposed to misinformation to believe in the accuracy of that false information without encountering any contradiction or correction [21]. However, the informational homogeneity of a certain sub-network may not emerge instantly, but instead increase over time: One study that focused on network fragmentation in the context of the Syrian war over a period of 32 months showed that fragmentation and homogeneity were generally high in the network. However, the temporal evolution of these fragmented groups showed that only one group increased its ideological homophily over time [65].

While some research has investigated the fragmentation process in political issues, there is still very little scientific understanding of fragmentation in the context of misinformation. An investigation on the online consumption of fake news found fragmentation between a fake news audience (minority) and a real news audience (majority) [66]. The same study also determined that the rapid spread of misinformation has a massive impact on the media environment, making it difficult for users to determine which news is right and which is wrong.

In addition to the existence of misinformation, however, the temporal consideration of informational homogeneity is particularly relevant in order to examine whether the dissemination of misinformation leads to the formation of disconnected network segments over time. In line with suggestions made by Webster & Ksiazek [62], we argue that audience fragmentation is best addressed by a network analytic approach, assessing the links between nodes in a communication network. To assess the fragmentation of the information landscape related to the COVID-19 pandemic, we therefore rely on the concept of informational homogeneity and its manifestation over time and ask:

RQ3: Are there temporal (i.e., monthly) differences in informational homogeneity within YouTube information networks in the context of the COVID-19 pandemic?

3. Methodology

In order to assess the proportion of misinformation, we first needed to: (a) collect data, (b) annotate part of the collected data, and (c) train a model to predict all remaining data records. This data consists of information about YouTube videos related to the search term "coronavirus," along with the comments on these videos. A random sample was then annotated by determining, for each video or comment, whether it belonged to the "misinformation" or "non-misinformation" class. Finally, natural language processing (NLP) techniques were used to predict the class for the remaining data records that had not been annotated. In particular, our approach uses the deep learning technique BERT (Bidirectional Encoder Representations from Transformers) to detect misinformation based on the previously annotated comments and videos on YouTube. To ensure the quality of the classification model, we performed an error analysis to ensure error classes and validate the results.

Once this classification step had been completed, to examine the communication network of YouTube and compute its informational homogeneity, we: (a) transformed the YouTube videos and comments into a network structure and (b) computed the external–internal (E-I) index on the basis of the two classes. This combination of NLP and network analysis allowed us to identify the homogeneity of the network from the communication paths of users. To determine the fragmentation, i.e., the temporal aspect of homogeneity in our data, we examined subsequent months individually and compared them with each other.

3.1. Data collection

For data collection, we ran a self-developed program from 1 January 2020 to 11 March 2020 that accesses the YouTube application programming interface (API) and retrieves metadata about the videos and content, as well as metadata of comments and replies. YouTube plays an increasingly important role in the consumption of news because it provides a platform where multifaceted information from different news channels comes together [67]. Based on a recent Pew Research Center survey, 26% of U.S. adults indicated that they used YouTube as a news source because it is a key source for staying up to date [68]. We used a similar method to that used by Röchert et al. [54] to obtain the data using the search, video, comments, and replies list. When passing the parameters responsible for the output of the search results, we sorted the videos with the parameter "order" by "date" in order to iteratively collect, for every single day, content related to the search term "coronavirus." By repeating this iterative procedure after short periods of time, it was possible to ensure that the number of collected videos could be heard. Furthermore, we carried out another collection in which we changed the parameter "order" to "relevance" in order to also collect the most relevant videos according to YouTube. For both procedures, we set the parameter "relevantLanguage" to "en" and "de" to get a wide range of videos. We searched for the word "coronavirus," which was used internationally at that time. Based on Google Trends and a worldwide comparison of the words "coronavirus" and "COVID-19," the term coronavirus received much higher attention during the investigation period[4]. Following data collection, we noticed that despite the filtering of the language, the term "coronavirus" was still used in multiple other languages. Focusing on the English language, we used the language classification API "detectlanguage"[5] to identify English videos based on the title and description. This step is necessary because, although we had specified a "relevance language" in our requests to YouTube's API, the API documentation warns that "results in other languages will still be returned if they are highly relevant to the search query term." In total, we collected 10,724 videos and 2,585,367 comments and replies. Figure 1 shows the crawling procedure of the dataset.

3.2. Annotation

We developed a coding scheme that serves as a guideline for the manual annotation of unlabeled videos and comments. For this purpose, we defined two mutually exclusive classes (misinformation and non-misinformation), which were used for annotating videos and comments. Misinformation is inaccurate information shared by the user without a clear intention to deceive. Often, the user is involved in circulating the misinformation without knowing the background truth, here in this study without knowing the truth about the YouTube videos. In contrast, disinformation is a piece of information that is deliberately misleading or biased. The user has the intention to mislead or deceive others. People alter the truth or repurpose the original story to spread propaganda, cheat people, etc. Without knowing the origin of YouTube videos, it is difficult to classify a video as misinformation or disinformation, so for this study, we classified videos as misinformation and non-misinformation. The misinformation category might include some videos that are disinformation, while in non-misinformation, we include YouTube videos that do not contain any false information.

In this study, the "misinformation" class contains all unintentionally

[4] https://trends.google.com/trends/explore?gprop=youtube&q=covid-19,%2Fm%2F01cpyy

[5] https://detectlanguage.com/

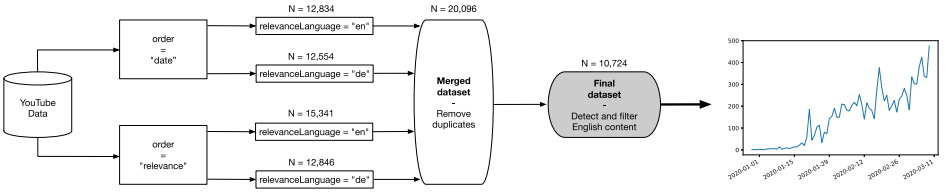

Figure 1. An illustration of the data collection process.

and intentionally false information about the origin, distribution, prevention, etc., of the COVID-19 virus and disease. This class also includes conspiracy theories and content that misleads the user with a wrong title, captions, misrepresented context, or statistics. In contrast, videos or comments that do not contain any information about the coronavirus or neutral news reporting, as well as satire or parodies, are annotated as "non-misinformation." Furthermore, this class includes videos or comments that do not contain any false information and therefore could be true or refer to a completely different topic. If the video or comment was not in English, it was also marked as non-misinformation. In line with ethical principles regarding misinformation that may lead to ostracism and profiling, we decided to consider only content-relevant information in the annotation; metadata such as the name of the channel was hidden or not considered in the video annotation.

To ensure the correct annotation, especially of the videos, the content was examined while watching the video and investigating the title and description, and the topic was additionally checked with the International Fact-Checking Network (IFCN) of the Poynter Institute. If the IFCN signatories had not fact-checked the information, then we searched for additional information from reliable sources such as government portals and reputable news websites.

Since annotating the entire dataset using this technique was not feasible, we annotated a portion that was sampled according to the number of videos and comments published in the respective months as follow: To ensure that all time periods were sufficiently represented in the sample, we used stratified sampling so that 20% of the sample consisted of data from January (when the overall number of videos about COVID-19 was still lower), 40% from February, and the remaining 40% from March. We only considered the videos that had public comments and found that some of the videos or comments had been deleted or removed from YouTube. The final sample consisted of 429 videos and 10,400 YouTube comments, which were annotated. An overview is presented in Table 1.

Each YouTube video and comment was annotated by three annotators, all undergraduate students. To measure inter-coder reliability, we used Fleiss' Kappa [65], which resulted in a value of 0.582 for the video dataset and a value of 0.473 for the comment dataset, indicating a moderate level of agreement. For the determination of the final class, we used a majority vote. If the class could not be determined, the annotators reviewed the videos and comments again in order to come to a decision.

Overall, the number of videos that contained misinformation was not sufficient to train a deep learning model. We pre-tested this in advance and found that the model overfitted due to the low training data and that too many errors occurred in the performance on the test data. This effect was not only observed with the undersampling procedure, but also with the distribution of the real dataset (unbalanced). As Zhang et al. [69] point out, a major challenge in developing a misinformation classification system is the lack of annotated data; therefore, we decided to add external data from the IFCN of the Poynter Institute, which stores known false information content about the coronavirus [41]. The database contains fact-checked articles on COVID-19 that have been identified from different signatories (fact-checking companies) from multiple countries. The IFCN provides basic information such as title, date, and country in English and points to the actual fact-checked article's webpage. A further advantage of these articles is that they cover a broad spectrum and report worldwide information regarding the COVID-19 pandemic, which is therefore ideal for the further course of our analysis. Since many of these statements are very short, they are similar to the YouTube video titles and are thus an ideal data source. Figure 2 demonstrates an example of the gathered information from Poynter.

For the collection of the data, we manually collected the headings from 14 January 2020 to 9 March 2020, which also corresponds to our investigation period and hence reflects comparable incidents related to the coronavirus. To obtain only clearly false information, we filtered the results to only include the category "false" (see Table 2).

Table 1
Overview of manually labeled YouTube videos and comments.

		Month			
Dataset	Class	January	February	March	**Total**
Videos	Misinformation	3	22	9	**34**
	Non-misinformation	61	152	182	**395**
Comments	Misinformation	119	379	298	**796**
	Non-misinformation	1283	3767	4554	**9604**

3.3. Pre-processing

As a first step in pre-processing the data, we merged the manually annotated video information with the fact-checked statements. In total, our video dataset contained 996 entries belonging to the misinformation class and 395 entries belonging to the non-misinformation class. For the comments, we used the 10,400 comments from the manual annotation. Since the class distributions were unbalanced in both datasets, we randomly undersampled the larger class so that both classes had the same size in the training process. As a result, we had 395 records for each class in the video dataset and 796 records in the comment dataset. Before training our classification model, we also performed common text pre-processing steps so that the text could be handled more efficiently by the algorithm. These processes were identical for both datasets. We removed the hyperlinks mentioned in the text and expanded contractions (e.g., "wasn't" to "was not", "we'll" to "we will"). We also removed punctuation marks from the text. Since we do not train our model on video files (video sequences), but only on the textual metadata given for the video, we decided to merge the title as well as the description of the YouTube videos to capture more meaning in the text and not lose essential information. Therefore, we concatenated the title and description together, while for the comments classification, we used only the textual information of the comments and replies from the videos.

3.4. Classification model

For the classification of misinformation in the comments and videos, we used the state-of-the-art neural network language model BERT, which has been pre-trained on a large corpus in order to solve language processing tasks [5]. An essential advantage of BERT is that it can be fine-tuned for task-specific datasets and allows high text classification accuracy even for smaller datasets. In the context of COVID-19, BERT

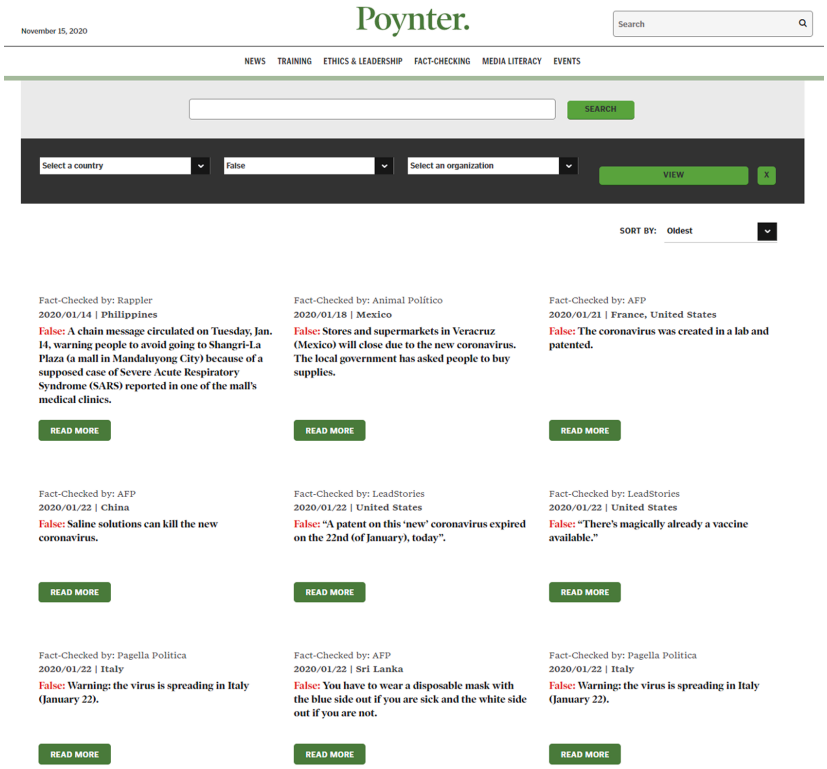

Figure 2. Screenshot of the Poynter database for COVID-19 in the category "false".

Table 2
Overview of additional fact-checked videos.

		Month			
Dataset	Class	January	February	March	**Total**
Fact-checked data	Misinformation	213	507	242	**962**

has already been applied for multiclass classification tasks, for example, on the Chinese social media platform Weibo, where it achieved considerable accuracy [70]. Furthermore, BERT was also used for other problems such as the detection of misinformation [71,72] or the identification of hate speech [73,74]. When using BERT, it should be noted that the texts must be formatted in a specific way in order to ensure that the training is carried out correctly. This pre-processing includes converting text to lowercase, tokenizing it, breaking words into word pieces, as well as attaching "CLS" and "SEP" tokens to represent the meaning of the entire sentence and to separate sentences for the next sentence prediction task. We split the video and comment data into 80% training data (videos: 632; comments: 1,273) and 20% test data (videos: 158; comments: 319). The randomization of the data prevents seasonal patterns from being learned by the model. For BERT fine-tuning, the model for the videos was trained for four epochs and the model for the comments was trained for three epochs with a learning rate of 2e-5. For the video dataset, we used a batch size of 8 with a sequence length of 128 because the dataset contains fewer records, and the titles of the fact-checking websites are also generally shorter. For the comments dataset, we chose a batch size of 32 with a sequence length of 128, since the average sequence length was 153 and the median sequence length was 88. After the individual prediction on the two test datasets, we evaluated the accuracy of the two models using the weighted F1 score.

3.4.1. *Baseline model*

We use two classical machine learning algorithms (Support Vector Machine [SVM] and Logistic Regression [LR]) and two deep learning techniques (Long Short-Term Memory [LSTM] and Convolutional Neural Network [CNN]) to compare the performance of BERT against those baseline models. For SVM and LR, we trained a term frequency–inverse document frequency (TF-IDF)-weighted character n-gram model with optimally selected hyperparameters based on grid search with five-fold cross-validation. The applied hyperparameters can be found in Appendix A.

In the deep learning techniques, we decided to keep the architecture the same for the comments and videos. For this reason, we will describe them globally, with individual parameters given in Appendix A.

In the LSTM model, our first layer is an embedded layer with an input length of 128. After this layer, an LSTM layer with 128 memory units is added. Following this layer, we set a dense layer with a unit of 128. The output layer is defined by one neuron with a sigmoid activation function. As an optimization function, we choose Adam [75] with the binary cross entropy loss function, suited for binary classification problems.

The CNN model is characterized by the first layer as an embedded layer with an input length of 128. After this layer, a Conv1D layer of 128 filters and a size of 3 with a ReLu activation function and max pooling of 3 is added. Following this layer, we set a flatten layer to reduce the dimension in our model and add a dense layer with a unit of 128. The output layer is the same as that described for the LSTM model with a single neuron and a sigmoid activation function.

After we had compared all the models, the results showed that BERT had the best performance in the video and in the comment dataset. The comparison of the different models and their performance can be seen in Table 3.

As can be seen in Table 3 above, the best F1 score for the commentary classifier was 0.81, and the score for the video classifier was 0.97. A detailed demonstration of the prediction within each class of the chosen BERT models can be found in Table 4. Since the values of the F1 score were acceptable for our further analysis, we proceeded to use the models to classify the entire dataset of videos and comments.

3.4.2. Error analysis

We performed an error analysis to evaluate the performance of the video and comment classification models. Therefore, we created an independent validation set that does not contain training and test sets and consists of 50 data records for each month of comment and video datasets. In total, we had 150 videos and 150 comments that we analyzed. Based on these sample datasets, we performed a manual analysis and checked the predicted content for their accuracy. In this manual analysis, the predicted values of the comments and videos were compared to the human annotation in which the comments were read and the videos were watched. The aim of the manual analysis is to identify specific classes of errors that may be responsible for the incorrect prediction and that have occurred most frequently. Since our models are binary classifiers, we can specifically address false negative and false positive errors.

Comments

Overall, we identified an 8% error rate of our 150 comments where these were predicted only as false positives. In diagnosing the predicted comments and their classes, we identified four reasons (off-topic, sarcasm/joke, lack of special knowledge, and lack of video context) that were responsible for the misclassification.

Off-topic: In this identified error class, which occurred most frequently, we could see that YouTube comments did not focus on the topic under investigation, "coronavirus," but rather dealt with different topics, which were kept very general.

Sarcasm/joke: This error class has already been found in other studies on hate speech and refers to comments that contain sarcastic or funny content. In particular, the topic of coronavirus was addressed here in conjunction with the eating habits and food that might have caused the disease (bats) and the treatment of the virus (handwashing).

Lack of special knowledge: We identified this error class because some misclassifications were related to healthcare information such as contagion, wearing masks, or information about the virus. This also includes information about specific locations that were not frequently included in the dataset.

Lack of video context: In this class of errors, we found errors that were directly related to the content of the YouTube video. For example, these comments contained spelling errors or declared the related video to be fake news.

Videos

As with the comments, we also manually checked a sample of the video dataset for errors. In general, we found an error rate of 7.33%, with false negative and false positive errors. Videos that were no longer available on the YouTube platform (N=20) were still coded based on the title and description to ensure comparability. In addition to the identified classes of errors, we noticed in particular that the descriptions had a major influence on the classification of the videos. While many official news channels add a description when publishing the videos, there are also some channels that do not have descriptions. Videos that do not have descriptions are more likely to be declared as misinformation by the algorithm. Overall, we were able to determine the following one class of error in the comments that were "false negative." The "conspiracy content" category was the most frequent with eight errors. In this category, as many as four videos had been deleted and were no longer available on YouTube due to violations of YouTube guidelines.

Conspiracy content: We defined this error class because it was most prevalent with conspiracy theory content about COVID. Here, the titles in particular consisted of rhetorical questions and were related to the outbreak of the virus. Furthermore, the length of the titles and the description of the videos were given with few characters.

For false positive errors, we were also able to identify one error class, in which the frequency of errors in the category "news channel content" occurred four times.

News channel content: The errors that were identified in this class were characterized by a short title in combination with a short description. More precisely, news channels used questions in the title (including rhetorical questions) and created a direct link to a specific scenario (e.g., disinfectant). Here, the description of the video may also be completely omitted.

3.5. Network analysis

After classifying the entire dataset of comments and videos using the trained models, we generated two different directed communication

Table 3
Model evaluation of deep learning and machine learning methods on the test dataset.

			Weighted average			Macro average		
Dataset	**Models**	**Epoch**	**Precision**	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
Comments		1	0.78	0.77	0.77	0.78	0.78	0.77
	BERT	2	0.80	0.80	0.80	0.80	0.80	0.80
		3	**0.81**	**0.81**	**0.81**	**0.81**	**0.81**	**0.81**
		4	0.81	0.81	0.81	0.81	0.81	0.81
	LSTM	10	0.71	0.70	0.70	0.71	0.71	0.70
	CNN	10	0.71	0.70	0.70	0.70	0.70	0.70
	LR	-	0.76	0.74	0.74	0.75	0.75	0.74
	SVM	-	0.75	0.74	0.74	0.74	0.74	0.74
Videos		1	0.97	0.97	0.97	0.97	0.97	0.97
	BERT	2	0.97	0.97	0.97	0.97	0.97	0.97
		3	0.97	0.97	0.97	0.97	0.97	0.97
		4	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**
	LSTM	10	0.92	0.91	0.91	0.91	0.92	0.91
	CNN	10	0.93	0.93	0.93	0.93	0.93	0.93
	LR	-	0.90	0.90	0.90	0.90	0.90	0.90
	SVM	-	0.91	0.91	0.91	0.90	0.91	0.90

Table 4
Summary of the precision, recall, and F1 for each class based on the final BERT models.

Dataset	Class	Metrics Precision	Recall	F1 score	Support	Prediction
Comments	Misinformation	0.79	0.81	0.80	149	154
	Non-misinformation	0.83	0.81	0.82	170	165
	Weighted avg.	0.81	0.81	**0.81**	319	319
Videos	Misinformation	0.97	0.96	0.97	72	71
	Non-misinformation	0.97	0.98	0.97	86	87
	Weighted avg.	0.97	0.97	**0.97**	158	158

networks (1. video comment, 2. comment network) from the entire predicted YouTube data. The distinction between the two networks is intended to clarify the analysis in terms of network homogeneity between video and comment misinformation.

The first network reflects the entire YouTube network with links to videos and comments. Here, the nodes represent uploaded videos and users who have written at least one comment. Interactions are represented by the directed edges. Nodes A and B are linked by a directed edge from A to B if: (a) user A has commented on video B or (b) user A has replied to a comment made by user B. The second network, on the other hand, was generated only from comments and their replies, in order to determine the communication within the comments. Videos that were represented previously as hubs were removed in this network.

Based on the output of the classification results for the videos and comments, we computed for each node whether the particular user has spread misinformation or not. In the case that users had written numerous comments, we computed the aggregated value of the classification outputs for each comment by applying the arithmetic mean (compare [54,80]). In addition, we also eliminated self-loops (comments regarding one's own video and replies to one's own comments) and disconnected nodes (videos without comments) because they have no further impact on the final outcome. To measure the informational homogeneity, we used the global E-I Index of Krackhardt & Stern [76]. The E-I index is defined as follows:

$$EI = \frac{E - I}{E + I}$$

where E represents the number of external ties and I the number of internal ties.

Furthermore, we computed the directed per-group E-I indices, considering the direction of the edges by counting only outgoing links as external links. In this context, the main purpose of the computed per-group E-I index is to focus on the interaction of the members of a specific group, i.e., which users they have communicated with. Compared to the undirected groupwise E-I index, this gives a much more accurate representation of users' interactions.

We performed a permutation test to determine whether the given E-I index is significantly smaller or larger than the expected E-I index when the connections in the network are randomly generated. This involves creating multiple iterations of graphs based on the sampling distribution, where each edge is randomly rewired. In this way, we can test the null hypothesis that the edges are randomly distributed among the nodes and ensure that the number of nodes in each group and the ties is constant.

Table 5 below provides an overview of the evaluated networks based on their network properties.

For a summary of our methodological approach, see Figure 3. First, the videos and their comments were collected using the YouTube API, and then a subset was manually annotated. We then trained the classifier on most of the annotated videos and comments using two independent BERT models and evaluated them using the remaining annotated data as test datasets. We then used the two trained models to classify the entire dataset and transformed the data into a network structure. Using this network, we were able to compute the informational homogeneity and determine how the discussion of misinformation developed over a period of three months.

4. Results

Regarding RQ1, we found that 26.37% (N = 681,811) of comments were classified as containing misinformation, while the proportion of non-misinformation content was 73.63% (N = 1,903,556). Of the videos, 3.5% (N = 376) contained misinformation and 96.5% (N = 10.348) non-misinformation. After aggregating the classifications across all of the content posted by each user, we found that in January, 16% of users primarily posted misinformation (compared with 84% who did not). In February, this number rose to 20%, and in March it dropped again to 16.4%. The proportion of misinformation from the interaction of videos and comments, which we could observe on the basis of our network perspective (after pre-processing), was 21.8% in January, 19.9% in February, and 16.3% in March. In order to validate the error of the classification and thus ensure the quality of the results, we decided to perform an error analysis. Based on this error analysis, we were able to identify four different error classes of the comments and two error classes of the videos, making a correct prediction of the comments difficult. The errors of the comments refer to thematic points of view, with a lack of additional information such as the content of the video watched or specific medical knowledge. Based on the content of the

Table 5
Network properties.

Network parameter	Video and comments network January	February	March	Comments networks January	February	March
Nodes	222,204	460,816	308,367	131,991	244,017	166,841
Edges	394,008	1,102,352	603,704	203,790	484,651	280,503
Avg. degree	1.77	2.39	1.96	1.54	1.99	1.68
Diameter	31	24	27	31	24	27
Max. out-degree	159	411	252	152	400	241
Max. in-degree	1,712	1,625	1,663	1,712	884	304
Density	0.000008	0.000005	0.000006	0.000012	0.000008	0.00001
Assortativity	0.004	0.005	0.019	0.062	0.060	0.067
Clustering coefficient	0.001	0.002	0.002	0.001	0.003	0.003

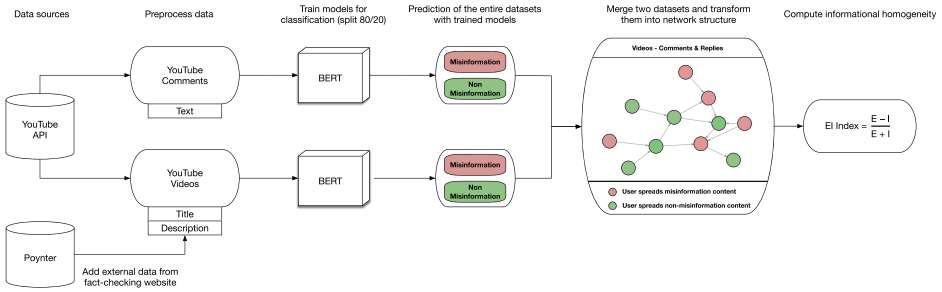

Figure 3. Description of the classification process and informational homogeneity analysis.

comments, we also found that comments that did not address the topic of coronavirus were misclassified and, thus, had a unusual number of words in the learned corpus, as well as containing sarcastic/funny content. On the other hand, when diagnosing the errors of the videos, we found that videos that have already been deleted, lack information, or contain conspiracy beliefs are also incorrectly predicted. Nevertheless, we have to mention that the percentage of errors in the analysis of errors is very low.

To address RQ2, the extent of informational homogeneity in YouTube networks among user-generated comments and videos on COVID-19, the results show that there is a significant difference in the class E-I indices of misinformation and non-misinformation. In our analyses of the two networks (video-comment network, comment network) over three months, the results indicate that people who disseminated misinformation find themselves in a heterogeneous discussion environment. Table 6 demonstrates the results of the homogeneity analysis of the three different months based on the video and comment network. While the per-class E-I indices for non-misinformation are all negative (January: –0.795, February: –0.850, March: –0.869) and thus show a homogeneous communication pattern, they are all positive for the class of misinformation (January: 0.788, February: 0.842, March: 0.839), which indicates heterogeneous communication. Similar results in Table 7 were also found in the communication-only network, where we considered only the links of comments and removed the links to the video. Compared to the whole network, the per-class E-I index of the commentary network has slightly lower values over the three months for all classes and therefore also a lower global E-I index. This results from the fact that videos, which are seen as a central hub in the network, are dropped and thus no longer have significant influence. Here, the per-class E-I indices for non-misinformation also show a homogeneity trend (January: –0.649, February: –0.698, March: –0.760), whereas the misinformation class indicates a more heterogeneous communication pattern (January: 0.508, February: 0.558, March: 0.605).

Taking into account the permutation test, the results in Table 6 and 7 indicate that the expected E-I index is negative for the "non-misinformation" class and positive for the "misinformation" class. With respect to the results on the null hypothesis test in Table 6, one can see that in all months the observed E-I index of the "non-misinformation" class is significantly closer to –1 than the expected E-I index, while the observed E-I index of the "misinformation" class is significantly closer to +1 than the expected E-I index. Concerning the null hypothesis test in the comment-only network, Table 7 shows that in all months the observed E-I index of the "non-misinformation" class is significantly closer to –1 than the expected E-I index. For the "misinformation" class, in contrast, one can see that in all months the expected E-I index of the "misinformation" class is significantly closer to +1 than the observed E-I index.

Addressing the RQ3, there is a trend in both networks (videos/comments, comments only) for communication to become more informationally homogeneous over time. A consideration of the global E-I index for both networks indicates a clear trend towards a more homogeneous information network over time. With respect to the individual classes, however, there are minor differences. While in the video comment network the values for the misinformation class become more heterogeneous from January to February, the E-I index stagnates at a similar value of 0.839 in March. In the pure commentary network, it can be seen that for the misinformation class the communication within the comments becomes continuously more heterogeneous from January to March. For the non-misinformation class, the findings show that the communication between the videos and the comments or only within the comments becomes more homogeneous from January to March.

5. Discussion

In the COVID-19 pandemic, the world has not only seen a virus spread all over the world—an overabundance of information, including misinformation and conspiracy theories, has also been disseminated through online social networks [77]. The fight against misinformation on social media platforms poses many challenges: One step towards addressing those challenges is to understand whether the diffusion of misinformation divides users into segments in online networks, leading some users—in the long run—to be caught in clusters that are predominantly filled with misinformation and disconnected from the clusters that provide corrections or contradictions. To examine this question, we used a combination of deep learning and network analysis methods to compute the informational homogeneity among videos and comments on COVID-19 on the video-sharing platform YouTube.

Results showed that, over the period from January to mid-March, approximately 3.5% of videos and 26.37% of comments contained misinformation. These findings of misleading videos are lower than the proportion found by Li et al. [19], who revealed that about 23%–26% of YouTube videos were misleading, generating attention from millions of viewers worldwide. A possible explanation for this might be that Li et al. analyzed data crawled on one day at the end of March 2020. This was undoubtedly a "hot" stage in which information needs might have been remarkably higher, but also the potential publication of misinformation in the form of videos may have likewise been higher. In our view, our results do not challenge the findings presented by Li et al., but indicate that the amount of misinformation may vary depending on the stage of a crisis. When comparing our results with those of Li et al., stages might be more ephemeral in the sense that the amount of information might not increase month by month (as we show in our results), but significantly from day to day. Future analyses need to investigate the emergence of misinformation in much smaller units to do justice to the information needs created in the face of (health) crises. Considering our results, we can see that the spread of videos containing misinformation is low and that some videos have already been deleted, but the number of comments containing misinformation and thus having an influence on users' information processing is relatively high at 26.37%. It seems even more

essential for social media service providers to take action against misinformation comments given that the spread of such comments could most certainly have severe consequences on individual and collective health [6,20].

Using error analysis with the validation dataset, we were able to examine the quality of the classification model and identify specific sources of error related to comments and videos. In the case of comments, it was noticeable that they were more often incorrectly predicted if, for example, they were not related to the context of COVID, and thus were off-topic, or if they required specific medical knowledge to correctly identify the context. In addition to these findings, however, there are also parallels with other research that has looked at text classification of hate speech on online social media, which also found sources of error from texts such as sarcasm [78,79]. Text classification seems to work better using state-of-the art techniques such as BERT, but errors still occur when there is ambiguity or too little context. Reviewing the videos, it was apparent that many videos had already ceased to exist, potentially having been deleted due to the current YouTube guidelines, as YouTube is increasingly taking action against misinformation.

Misinformation in the domain of public health can pose a significant risk if people believe in the accuracy of this information and act accordingly. The mistaken belief in the accuracy of misinformation could be reinforced if that misinformation is embedded in a network in which misinformation is predominantly present without any correction or contradiction [21]. To analyze the networks in which misinformation is spread, we transformed our YouTube dataset into a network and computed the extent to which this discussion network may contain homogeneous clusters. Our results indicate that the communication paths of users who disseminate misinformation in the network are quite heterogeneous, since they are predominantly connected with nodes that disseminate non-misinformation. The E-I index indicated a relatively high level of informational heterogeneity associated with misinformation and this pattern slightly increased over time, suggesting that the spread of misinformation does not lead to an increase in *misinformational* homogeneity in networks in the long run. This result would speak against the notion of network fragmentation consisting of enclaves with certain types of information that are not available to others [59,62].

In this context, it seems worthwhile to compare the level of informational homogeneity between networks containing videos and comments versus networks containing only comments (see Tables 6 and 7): In fact, results showed that the misinformation was connected to non-misinformation to a larger extent when networks included both types of content, i.e., videos and user-generated comments. Therefore, it seems that the blending of mass and interpersonal communication that characterizes many social media platforms [48] is responsible for higher levels of informational heterogeneity. While this appears to be a desirable result, it also raises questions: Given that the prevalence of COVID-19-related misinformation was higher in user-generated comments than in videos, future (experimental) research needs to test under which circumstances user-generated comments challenging or contradicting health-related information featured in journalistic videos or articles can exert an impact on their viewers'/readers' ultimate health-related knowledge and attitudes (e.g., on the acceptance of a COVID-19 vaccine).

Table 6
Results of the permutation test with the observed and expected class E-I index (videos and comments network).

Month	Sentiment	Observed E-I index	Expected E-I index	P (obs ≥exp)	P (obs ≤ exp)
January	Global	–0.508	–0.318	<0.01*	1.00
	Misinformation	0.788	0.564	1.00	<0.01*
	Non-misinformation	–0.795	–0.564	<0.01*	1.00
February	Global	–0.587	–0.363	<0.01*	1.00
	Misinformation	0.842	0.603	1.00	<0.01*
	Non-misinformation	–0.850	–0.603	<0.01*	1.00
March	Global	–0.653	–0.455	<0.01*	1.00
	Misinformation	0.839	0.674	1.00	<0.01*
	Non-misinformation	–0.869	–0.674	<0.01*	1.00

Table 7
Results of the permutation test with the observed and expected class E-I index (comment-only network).

Month	Sentiment	Observed E-I index	Expected E-I index	P (obs ≥ exp)	P (obs ≤ exp)
January	Global	–0.488	–0.383	<0.01*	1.00
	Misinformation	0.508	0.619	<0.01*	1.00
	Non-misinformation	–0.649	–0.619	<0.01*	1.00
February	Global	–0.553	–0.405	<0.01*	1.00
	Misinformation	0.558	0.637	<0.01*	1.00
	Non-misinformation	–0.698	–0.637	<0.01*	1.00
March	Global	–0.632	–0.491	<0.01*	1.00
	Misinformation	0.605	0.701	<0.01*	1.00
	Non-misinformation	–0.760	–0.701	<0.01*	1.00

There are two possible interpretations of our results, one optimistic, one pessimistic, yet both equally valid. The fact that misinformation is not concentrated in closed networks consisting of nodes that are predominantly associated with false information may prevent the formation of cohesive groups in which individuals mutually reinforce misperceptions and attitudes [64]. At the same time, it seems that misinformation successfully diffused in mainstream networks that were otherwise filled with non-misinformation. While this certainly does not lead to a segregation of certain information consumers, it may make the detection of misinformation more difficult for users who encounter false information in juxtaposition with accurate information [37]. At this point, it remains unclear whether misinformation is spread deliberately in those networks.

6. Limitations

As with most research, this research also has a number of limitations. First, we would like to emphasize that our results are based only on an English language dataset and on one specific search keyword, "coronavirus." Thus, we cannot state whether the results are transferable to other languages. Due to this random factor of sampling, we were faced with the challenge that there were too few datasets in the video dataset for the training of the BERT model, and we overcame this by increasing the amount of under-represented data by using fact-checking. In addition, time passed during the data collection and annotation process, which led to some videos being removed from YouTube due to violations of the guidelines and, thus, also excluded from our data analysis. Another limitation is the fact that we analyzed content published at the beginning of the pandemic; more precisely, we analyzed the videos and comments from 1 January 2020 to 11 March 2020. For this reason, it should be pointed out that after this period of time, further videos as well as comments may have been produced, thereby potentially providing more misinformation. For this reason, we cannot make any statements about the further course of the pandemic. A more comprehensive analysis could include later months of the pandemic and cover the full information landscape related to COVID-19 on YouTube. Moreover, it is worthy of note that our conclusions are based on predictions by a deep learning model (BERT), which has shown good results in previous research in different areas. The results should nevertheless be considered with some circumspection, since our results show that despite the high F1 score, there are still a few incorrect classifications in the test dataset. The final limitation is that YouTube Data API developer policy does not allow publication or distribution of the data used in this study, which does not ensure reproduction of the same results.

7. Conclusion and future work

This study investigated the informational homogeneity of misinformation on YouTube in the context of the current COVID-19 pandemic. We annotated random comments and videos from YouTube between January and March that were relevant to the search keyword "coronavirus" and applied a combination of NLP and network analysis to compute the informational homogeneity. The results showed that, despite small variations regarding the proportion of misinformation on YouTube between the three months analyzed, approximately one third of the content contained certain forms of misinformation. One of the more significant findings of this study is that although misinformation exists on YouTube, it is not concentrated in homogeneous networks filled with predominantly false information—instead, misinformation is moderately associated with non-misinformation. This finding indicates that the YouTube network is not fragmented in the sense that some groups are largely confronted with misinformation while others are not. Since our analysis is limited to the keyword "coronavirus," it would also be interesting for future research to include keywords that are explicitly related to misinformation or conspiracy theories. Thus, network structures based on single conspiracy theories could be investigated to get an even more precise understanding of (mis)informational homogeneity in online networks. Future work may also involve using additional metadata from videos (i.e., visual, audio and subtitles) to improve the automatic classification of misinformation. Also, it would be worthwhile to investigate the spread of misinformation and the identification of relevant actors in the network with their intentions. Our findings could be complemented by analyses of regional differences in the spread of misinformation, to examine whether users in some parts of the world are more likely to receive misinformation on a public health crisis. Addressing these questions could help to assess the actual role of social media platforms in shaping information diffusion processes and fostering the spread of misinformation that could put global health at risk.

CRediT authorship contribution statement

Daniel Röchert: Data curation, Investigation, Conceptualization, Methodology, Software, Validation, Formal analysis, Project administration, Writing – original draft, Writing – review & editing. **Gautam Kishore Shahi:** Data curation, Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **German Neubaum:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing, Funding acquisition. **Björn Ross:** Methodology, Writing – review & editing. **Stefan Stieglitz:** Conceptualization, 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.

Acknowledgements

This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia (Grant Number: 005-1709-0004), Junior Research Group "Digital Citizenship in Network Technologies" (Project Number: 1706dgn009).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.osnem.2021.100164.

References

[1] N. Newman, R. Fletcher, A. Schulz, S. Andi, R. Nielsen, Reuter Institute for the Study of Journalism. Digital news report 2020, 2020 https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf.
[2] H. Ding, J. Zhang, Social media and participatory risk communication during the H1N1 flu epidemic: a comparative study of the United States and China, China Media Research 6 (2010) 80–91. https://scholar.google.com/scholar_lookup?hl=en&volume=6&publication_year=2010&pages=80-90&journal=China+Media+Res&author=Ding+H.&author=Zhang+J.&title=Social+media+and+participatory+risk+communication+during+the+H1N1+flu+epidemic%3A+a+comparative+study+of+the+United+States+and+China.
[3] C. Huang, X. Xu, Y. Cai, Q. Ge, G. Zeng, X. Li, W. Zhang, C. Ji, L. Yang, Mining the characteristics of COVID-19 patients in China: analysis of social media posts, J Med Internet Res 22 (2020) e19087, https://doi.org/10.2196/19087.
[4] Y. Wang, M. McKee, A. Torbica, D. Stuckler, Systematic literature review on the spread of health-related misinformation on social media, Social Science & Medicine 240 (2019), 112552, https://doi.org/10.1016/j.socscimed.2019.112552.
[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, ArXiv Preprint ArXiv:1810.04805. (2018).
[6] N.M. Krause, I. Freiling, B. Beets, D. Brossard, Fact-checking as risk communication: the multi-layered risk of misinformation in times of COVID-19, Journal of Risk Research 0 (2020) 1–8, https://doi.org/10.1080/13669877.2020.1756385.
[7] A. Mitchell, M. Jurkowitz, J. Oliphant, E. Shearer, Three months in, many Americans see exaggeration, conspiracy theories, and partisanship in COVID-19 news, Pew Research Center (2020).
[8] W.H. Organization, Novel Coronavirus (2019-nCoV): situation report, 13, World Health Organization, 2020.
[9] R.S. D'Souza, S. D'Souza, N. Strand, A. Anderson, M.N.P. Vogt, O. Olatoye, YouTube as a source of medical information on the novel coronavirus 2019 disease (COVID-19) pandemic, Global Public Health 15 (2020) 935–942, https://doi.org/10.1080/17441692.2020.1761426.
[10] D.A. Scheufele, N.M. Krause, I. Freiling, D. Brossard, How not to lose the COVID-19 communication war, Issues in Science and Technology 17 (2020). https://issues.org/covid-19-communication-war/.
[11] R. Pathak, D.R. Poudel, P. Karmacharya, A. Pathak, M.R. Aryal, M. Mahmood, A. A. Donato, YouTube as a source of information on Ebola virus disease, North American Journal of Medical Sciences 7 (2015) 306.
[12] K. Bora, D. Das, B. Barman, P. Borah, Are internet videos useful sources of information during global public health emergencies? a case study of YouTube videos during the 2015–16 Zika virus pandemic, Pathogens and Global Health 112 (2018) 320–328.
[13] A. Pandey, N. Patni, M. Singh, A. Sood, G. Singh, YouTube as a source of information on the H1N1 influenza pandemic, American Journal of Preventive Medicine 38 (2010) e1–e3.
[14] A. Gruzd, P. Mai, Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter, Big Data & Society 7 (2020), 205395172093840, https://doi.org/10.1177/2053951720938405.
[15] G. Kawchuk, J. Hartvigsen, S. Harsted, C.G. Nim, L. Nyirö, Misinformation about spinal manipulation and boosting immunity: an analysis of Twitter activity during the COVID-19 crisis, Chiropr Man Therap 28 (2020) 34, https://doi.org/10.1186/s12998-020-00319-4.
[16] R. Kouzy, J. Abi Jaoude, A. Kraitem, M.B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E.W. Akl, K. Baddour, Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter, Cureus (2020) 12, https://doi.org/10.7759/cureus.7255.
[17] G.K. Shahi, A. Dirkson, T.A. Majchrzak, An exploratory study of COVID-19 misinformation on Twitter, Online Social Networks and Media 22 (2021) 100104, https://doi.org/10.1016/j.osnem.2020.100104.
[18] P. Mena, D. Barbe, S. Chan-Olmsted, Misinformation on Instagram: the impact of trusted endorsements on message credibility, Social Media+ Society 6 (2020), https://doi.org/10.1177/2056305120935102, 2056305120935102.
[19] H.O.-Y. Li, A. Bailey, D. Huynh, J. Chan, YouTube as a source of information on COVID-19: a pandemic of misinformation? BMJ Glob Health 5 (2020), e002604 https://doi.org/10.1136/bmjgh-2020-002604.
[20] S. Tasnim, M.M. Hossain, H. Mazumder, Impact of rumors and misinformation on COVID-19 in social media, J Prev Med Public Health 53 (2020) 171–174, https://doi.org/10.3961/jpmph.20.094.
[21] D.A. Scheufele, N.M. Krause, Science audiences, misinformation, and fake news, Proc Natl Acad Sci USA. 116 (2019) 7662–7669, https://doi.org/10.1073/pnas.1805871115.
[22] J. Bright, Explaining the emergence of political fragmentation on social media: the role of ideology and extremism, Journal of Computer-Mediated Communication 23 (2018) 17–33, https://doi.org/10.1093/jcmc/zmx002.
[23] C.R. Sunstein, #Republic: divided democracy in the age of social media, Princeton University Press, Princeton ; Oxford, 2017.
[24] W.H. Organization, Weekly Operational Update on COVID-19 - 6 September 2021. https://www.who.int/docs/default-source/coronaviruse/weekly-updates/wou_20-nov_cleared.pdf.
[25] Y. Kryvasheyeu, H. Chen, N. Obradovich, E. Moro, P. Van Hentenryck, J. Fowler, M. Cebrian, Rapid assessment of disaster damage using social media activity, Science Advances 2 (2016), e1500779.

[26] J. Li, H.R. Rao, Twitter as a rapid response news service: an exploration in the context of the 2008 China earthquake, The Electronic Journal of Information Systems in Developing Countries 42 (2010) 1–22.
[27] A. Goel, L. Gupta, Social Media in the Times of COVID-19, Journal of Clinical Rheumatology : Practical Reports on Rheumatic & Musculoskeletal Diseases. 26 (2020) 220–223. https://doi.org/10.1097/RHU.0000000000001508.
[28] S. Stieglitz, D. Bunker, M. Mirbabaie, C. Ehnis, Sense-making in social media during extreme events, J Contingencies Crisis Man 26 (2018) 4–15, https://doi.org/10.1111/1468-5973.12193.
[29] M. Mirbabaie, D. Bunker, S. Stieglitz, J. Marx, C. Ehnis, Social media in times of crisis: learning from Hurricane Harvey for the coronavirus disease 2019 pandemic response, Journal of Information Technology 35 (3) (2020), 026839622092925, https://doi.org/10.1177/0268396220929258.
[30] A. Mitchell, J. Oliphant, Americans immersed in COVID-19 news; most think media are doing fairly well covering it, Pew Research Center 18 (2020).
[31] H. Allcott, M. Gentzkow, C. Yu, Trends in the diffusion of misinformation on social media, Research & Politics 6 (2019), 205316801984855, https://doi.org/10.1177/2053168019848554.
[32] R. Ehrenberg, Social media sway: Worries over political misinformation on Twitter attract scientists' attention, Science News 182 (2012) 22–25, https://doi.org/10.1002/scin.5591820826.
[33] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151, https://doi.org/10.1126/science.aap9559.
[34] B. Nyhan, J. Reifler, When corrections fail: the persistence of political misperceptions, Polit Behav 32 (2010) 303–330, https://doi.org/10.1007/s11109-010-9112-2.
[35] J. Shin, L. Jian, K. Driscoll, F. Bar, The diffusion of misinformation on social media: Temporal pattern, message, and source, Computers in Human Behavior 83 (2018) 278–287, https://doi.org/10.1016/j.chb.2018.02.008.
[36] D. Jolley, K.M. Douglas, The effects of anti-vaccine conspiracy theories on vaccination intentions, PloS One 9 (2014) e89177.
[37] L. Bode, E.K. Vraga, See something, say something: correction of global health misinformation on social media, Health Communication 33 (2018) 1131–1140, https://doi.org/10.1080/10410236.2017.1331312.
[38] M.J. Wood, K.M. Douglas, R.M. Sutton, Dead and alive: beliefs in contradictory conspiracy theories, Social Psychological and Personality Science 3 (2012) 767–773, https://doi.org/10.1177/1948550611434786.
[39] A. Bessi, M. Coletto, G.A. Davidescu, A. Scala, G. Caldarelli, W. Quattrociocchi, Science vs conspiracy: collective narratives in the age of misinformation, PloS One 10 (2015), e0118093.
[40] G.K. Shahi, D. Nandini, FakeCovid-A multilingual cross-domain fact check news dataset for COVID-19. Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media (2020), in: http://workshop-proceedings.icwsm.org/pdf/2020_14.pdf.
[41] G.K. Shahi , T.A Majchrzak , AMUSED: An Annotation Framework of Multi-modal Social Media Data, ArXiv:2010.00502 [Cs]. (2020). http://arxiv.org/abs/2010.00502 (accessed March 7, 2021).
[42] J.S. Brennen, F. Simon, P.N. Howard, R.K. Nielsen, Types, sources, and claims of Covid-19 misinformation, Reuters Institute 7 (2020).
[43] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C.M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, A. Scala, The COVID-19 social media infodemic, Sci Rep 10 (2020) 16598, https://doi.org/10.1038/s41598-020-73510-5.
[44] J.C.M. Serrano, O. Papakyriakopoulos, S. Hegelich, NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube, in: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, 2020.
[45] S. Taylor, B. Pickering, P. Grace, M. Boniface, V. Bakir, danah boyd, S. Engesser, R. Epstein, N. Fawzi, P. Fernbach, D. Fisher, B.G. Gardner, K. Jacobs, S. Jacobson, B. Krämer, A. Kucharski, A. McStay, H. Mercier, M. Metzger, F. Polletta, W. Quattrociocchi, S. Sloman, D. Sperber, C.H.B.M. Spierings, C. Wardle, F. Zollo, A. Zubiaga, Opinion forming in the digital age, Zenodo (2018), https://doi.org/10.5281/ZENODO.1468575.
[46] J. Tucker, A. Guess, P. Barbera, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, B. Nyhan, Social media, political polarization, and political disinformation: a review of the scientific literature, SSRN Journal (2018), https://doi.org/10.2139/ssrn.3144139.
[47] B.E. Weeks, H. Gil de Zúñiga, Six observations for the future of political misinformation research, American Behavioral Scientist (2019), 0002764219878236, https://doi.org/10.1177/0002764219878236.
[48] G. Neubaum, N.C. Krämer, Opinion climates in social media: blending mass and interpersonal communication: opinion climates in social media, Hum Commun Res 43 (2017) 464–476, https://doi.org/10.1111/hcre.12118.
[49] J. Shin, K. Thorson, Partisan selective sharing: the biased diffusion of fact-checking messages on social media: sharing fact-checking messages on social media, J Commun 67 (2017) 233–255, https://doi.org/10.1111/jcom.12284.
[50] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: homophily in social networks, Annual Review of Sociology 27 (2001) 415–444, https://doi.org/10.1146/annurev.soc.27.1.415.
[51] E. Bakshy, S. Messing, L.A. Adamic, Exposure to ideologically diverse news and opinion on Facebook, Science 348 (2015) 1130–1132, https://doi.org/10.1126/science.aaa1160.
[52] A. Boutyline, R. Willer, The social structure of political echo chambers: variation in ideological homophily in online networks, Political Psychology 38 (2017) 551–569.
[53] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H.E. Stanley, W. Quattrociocchi, The spreading of misinformation online, Proceedings of the National Academy of Sciences 113 (2016) 554–559, https://doi.org/10.1073/pnas.1517441113.
[54] D. Röchert, G. Neubaum, B. Ross, F. Brachten, S. Stieglitz, Opinion-based homogeneity on YouTube: combining sentiment and social network analysis, Computational Communication Research 2 (2020) 81–108, https://doi.org/10.5117/CCR2020.1.004.ROCH.
[55] L. Tang, K. Fujimoto, M. (Tuan) Amith, R. Cunningham, R.A. Costantini, F. York, G. Xiong, J.A. Boom, C. Tao, Down the rabbit hole" of vaccine misinformation on YouTube: network exposure study, J Med Internet Res 23 (2021) e23262, https://doi.org/10.2196/23262.
[56] E. Hussein, P. Juneja, T. Mitra, Measuring misinformation in video search platforms: an audit study on YouTube, Proc. ACM Hum.-Comput. Interact. 4 (2020) 1–27, https://doi.org/10.1145/3392854.
[57] J. Shin, L. Jian, K. Driscoll, F. Bar, Political rumoring on Twitter during the 2012 US presidential election: rumor diffusion and correction, New Media & Society 19 (2017) 1214–1235, https://doi.org/10.1177/1461444816634054.
[58] R. Fletcher, R.K. Nielsen, Are News Audiences Increasingly Fragmented? A Cross-National Comparative Analysis of Cross-Platform News Audience Fragmentation and Duplication, Journal of Communication 67 (4) (2017) 476–498, https://doi.org/10.1111/jcom.12315.
[59] S. Flaxman, S. Goel, J.M. Rao, Filter Bubbles, Echo Chambers, and Online News Consumption, PUBOPQ. 80 (2016) 298–320. https://doi.org/10.1093/poq/nfw006.
[60] P. Mancini, Media Fragmentation, Party System, and Democracy, The International Journal of Press/Politics. 18 (2013) 43–60. https://doi.org/10.1177/1940161212458200.
[61] B.E. Weeks, T.B. Ksiazek, R.L. Holbert, Partisan enclaves or shared media experiences? a network approach to understanding citizens' political news environments, Journal of Broadcasting & Electronic Media 60 (2016) 248–268, https://doi.org/10.1080/08838151.2016.1164170.
[62] J.G. Webster, T.B. Ksiazek, The dynamics of audience fragmentation: public attention in an age of digital media, Journal of Communication 62 (2012) 39–56, https://doi.org/10.1111/j.1460-2466.2011.01616.x.
[63] T. Harper, The big data public and its problems: Big data and the structural transformation of the public sphere, New Media & Society 19 (2017) 1424–1439, https://doi.org/10.1177/1461444816642167.
[64] A. Sîrbu, D. Pedreschi, F. Giannotti, J. Kertész, Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model, PLoS ONE 14 (2019), e0213246, https://doi.org/10.1371/journal.pone.0213246.
[65] D. Freelon, M. Lynch, S. Aday, Online fragmentation in wartime: a longitudinal analysis of tweets about Syria, 2011–2013, The ANNALS of the American Academy of Political and Social Science 659 (2015) 166–179.
[66] J.L. Nelson, H. Taneja, The small, disloyal fake news audience: the role of audience availability in fake news consumption, New Media & Society 20 (2018) 3720–3737, https://doi.org/10.1177/1461444818758715.
[67] J.M. Sumiala, M. Tikka, Broadcast yourself-global news! a netnography of the "Flotilla" news on YouTube: broadcast yourself-global news!, communication, Culture & Critique 6 (2013) 318–335, https://doi.org/10.1111/cccr.12008.
[68] G. Stocking, P. van Kessel, M. Barthel, K.E. Matsa, M. Khuzam, Many Americans get news on YouTube, where news organizations and independent producers thrive side by side, Pew Research Centre (2020).
[69] D. Zhang, L. Zhou, J. Lim, From networking to mitigation: the role of social media and analytics in combating the COVID-19 pandemic, Information Systems Management (2020) 1–9, https://doi.org/10.1080/10580530.2020.1820635.
[70] T. Wang, K. Lu, K.P. Chow, Q. Zhu, COVID-19 sensing: negative sentiment analysis on social media in China via BERT model, IEEE Access 8 (2020) 138162–138169, https://doi.org/10.1109/ACCESS.2020.3012595.
[71] Y. Geng, Z. Lin, P. Fu, W. Wang, Rumor detection on social media: a multi-view model using self-attention mechanism, in: J.M.F. Rodrigues, P.J.S. Cardoso, J. Monteiro, R. Lam, V.V. Krzhizhanovskaya, M.H. Lees, J.J. Dongarra, P.M. A. Sloot (Eds.), Computational Science – ICCS 2019, Springer International Publishing, Cham, 2019, pp. 339–352, https://doi.org/10.1007/978-3-030-22734-0_25.
[72] E. Masciari, V. Moscato, A. Picariello, G. Sperli, A deep learning approach to fake news detection, in: D. Helic, G. Leitner, M. Stettinger, A. Felfernig, Z.W. Raś (Eds.), Foundations of Intelligent Systems, Springer International Publishing, Cham, 2020, pp. 113–122.
[73] K. Florio, V. Basile, M. Polignano, P. Basile, V. Patti, Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media, Applied Sciences. 10 (2020) 4180. https://doi.org/10.3390/app10124180.
[74] J. Pavlopoulos, N. Thain, L. Dixon, I. Androutsopoulos, Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 571–576.
[75] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, ArXiv Preprint ArXiv:1412.6980. (2014).

[76] D. Krackhardt, R.N. Stern, Informal Networks and Organizational Crises: An Experimental Simulation, Social Psychology Quarterly 51 (1988) 123–140, https://doi.org/10.2307/2786835.
[77] W.H. Organization, Naming the coronavirus disease (COVID-19) and the virus that causes it, (2020). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it.
[78] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 145–153.
[79] B. van Aken, J. Risch, R. Krestel, A. Löser, Challenges for toxic comment classification: An in-depth error analysis, ArXiv Preprint ArXiv:1809.07572. (2018).
[80] D. Röchert, G. Neubaum, B. Ross, S. Stieglitz, Caught in a networked collusion? homogeneity in conspiracy-related discussion networks on YouTube, Information Systems (2021), 101866, https://doi.org/10.1016/j.is.2021.101866.

Research Paper 5: "The homogeneity of right-wing populist and radical content in YouTube recommendations"

Type	Conference
Rights and permission	Re-used from ACM
Authors	**Röchert, Daniel**; Weitzel, Muriel; Ross, Björn
Year	2020
Outlet	Social Media and Society
Publisher	Association for Computing Machinery
Permalink/DOI	https://doi.org/10.1145/3400806.3400835
Full citation	Röchert, D., Weitzel, M., & Ross, B. (2020). The homogeneity of right-wing populist and radical content in YouTube recommendations. In *International Conference on Social Media and Society* (pp. 245-254).

The homogeneity of right-wing populist and radical content in YouTube recommendations

Daniel Röchert
University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science, Duisburg, Germany
daniel.roechert@uni-due.de

Muriel Weitzel
University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science, Duisburg, Germany
muriel.weitzel@stud.uni-due.de

Björn Ross
University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science, Duisburg, Germany
bjoern.ross@uni-due.de

ABSTRACT

The use of social media to disseminate extreme political content on the web, especially right-wing populist propaganda, is no longer a rarity in today's life. Recommendation systems of social platforms, which provide personalized filtering of content, can contribute to users forming homogeneous cocoons around themselves. This study investigates YouTube's recommendations system based on 1,663 German political videos in order to analyze the homogeneity of the related content. After examining two datasets (right-wing populist and politically neutral videos), each consisting of ten initial videos and their first and second level recommendations, we show that there is a high degree of homogeneity of right-wing populist and neutral political content in the recommendation network. These findings offer preliminary evidence on the role of YouTube recommendations in fueling the creation of ideologically like-minded information spaces.

CCS CONCEPTS

• **Information systems**; • **World Wide Web**; • **Web searching and information discovery**; • **Social recommendation**;

KEYWORDS

Filter bubble, Network Analysis, YouTube, right-wing populism

ACM Reference Format:
Daniel Röchert, Muriel Weitzel, and Björn Ross. 2020. The homogeneity of right-wing populist and radical content in YouTube recommendations. In *International Conference on Social Media and Society (SMSociety '20), July 22–24, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3400806.3400835

1 INTRODUCTION

In view of the constantly growing amount of user data, social media platforms provide personalized information for each user and filter it according to his or her individual characteristics. This can also mean that users are mostly or always shown content that corresponds to their interests and opinions. Especially when it comes to gauging opinion trends, social media are often used to serve as the voice of the people in populist movements [12]. Literature has proposed that this filtered consumption of information online could result in distorted perceptions of public opinion [25] and even the ideological fragmentation of society [37].

The challenge is to gain a deeper understanding of the transparency of these recommendation systems. Current research disagrees on the existence of filter bubbles caused by recommendation systems. While some studies find evidence and argue for the existence of filter bubbles on social platforms [1, 26], there are also empirical works speaking against the existence of ideologically one-sided filter bubbles in online media [6, 14, 40].

The video platform YouTube is one of the most frequently used social media sites. According to a recent study, it is used by 73 percent of Americans [30]. On YouTube, users are shown a list of recommended videos next to the video they are watching. This list is generated by an algorithm whose details are proprietary and therefore not published. In a recent paper [38], Google employees describe their two key objectives as engagement and satisfaction, where engagement is measured as the time a user spends watching recommended videos. In other words, YouTube's recommendation system is designed to users keep watching more videos. When a user is watching one of the recommended videos, this video comes with its own recommendations. We refer to these as second level recommendations.

In the light of this, it seems plausible that YouTube's recommendation system might constitute a filter bubble. Users consume political news on YouTube, and unlike in traditional journalism, a balanced presentation of different political viewpoints is not among the goals of the algorithm's developers. It is therefore necessary to question the extent to which these algorithms connect politically neutral and extreme videos and whether the induced recommendation networks are ideologically homogeneous or heterogeneous. No previous research, to our knowledge, has compared recommendations of right-wing populist and politically neutral videos on YouTube and examined their ideological homogeneity with network analysis techniques such as the E-I index.

Since there are already various studies on recommendation systems, a detailed analysis is required, which should include (a) an assessment of the political context of the videos and (b) a calculation of the homophily of the content in relation to the associated content. This work investigates how right-wing populist videos and politically neutral news videos are connected by the YouTube recommendation algorithm and the extent to which this leads to the creation of homogeneous networks. We applied network analysis

SMSociety '20, July 22–24, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7688-4/20/07…$15.00
https://doi.org/10.1145/3400806.3400835

techniques to two datasets in order to ask how the recommendations at the first and second levels differ from the initial videos in their content (right-wing populist or neutral).

2 THEORY

2.1 Personalized recommendation algorithms and their effect

Personalized algorithms on social media platforms aim to automatically provide users with personalized content that they are most likely to consume based on search history, click behavior or current location. This means that information that is not recommended by the algorithm is not displayed to the user. According to the "filter bubble" theory [29], social media users are more likely to be confronted with homogeneous content that is consistent with their opinions and values than with cross-cutting opinions. In other words, information that is not in line with one's point of view is filtered out of one's information feed, effectively capturing users in closed and like-minded "filter bubbles". Particularly when political and ideological opinions in the network are concerned, these filter bubbles can be seen as a threat to democracies through their selection of information, since people are only provided with information that corresponds to their interests or supports their current opinion [40], narrowing the political horizon of the discussion. Additionally, personalized filtering can limit users' autonomy and information choice [4].

The potential existence of filter bubbles and recommendation systems has already been investigated on various platforms such as Facebook, YouTube or Google Search. Bakshy et al. [3] analyzed 10 million Facebook newsfeeds from US citizens who reported political affiliations in their profiles to investigate the ideological diversity of news and opinions. Their findings showed that the Facebook algorithm is less likely to display cross-cutting content, leaving the users themselves primarily responsible, since they select content that corresponds to their political beliefs. The video platform YouTube with its recommendation system for related videos serves as the main mechanism of encouraging users to view new videos and discover new content [11, 39]. Indeed, 81 percent of Americans occasionally watch videos suggested by the platform's recommendation algorithm [36]. Furthermore, the YouTube recommendation system is mainly responsible for users being exposed to new content and staying on YouTube for longer periods of time. Findings have revealed that after the first recommended video, the average length of a video increased by around three minutes, while at a recommendation level of five (i.e., five successive recommendations), the average length of a video was 15 minutes more than the original video.

In addition to general results about the recommendation system on YouTube, some studies also focus on the recommendation of videos that espouse extreme political views. A recent study [21] used around 800 YouTube channels to investigate whether the recommendation algorithm has an impact on users and suggests more radicalized content to them. According to their results, radicalized users are not encouraged by the algorithm to encounter more extreme content. Conversely, O'Callaghan et al. [26] analyzed English and German-language extreme right-wing YouTube channels that are propagated by extreme right-wing Twitter accounts. Their results showed that users who click on a right-wing extremist video are very likely to get recommendations for further right-wing extremist videos.

Ottoni et al. [27] performed a content analysis to identify characteristics in YouTube captions and YouTube comments on alt-right channels based on one specific channel ("The Alex Jones Channel") and his 12 featured channels. According to their results, these channels used words such as "war", "terrorism" or "bombing" more often in their titles than those in a baseline dataset, while comments on these videos were more likely to refer to words like "Ebola", "radiation" and "virus". In general, these channels used words related to negative feelings more often than the study's baseline channels, which tended to use more positively associated words. This might be especially dangerous for the consumption of right-wing populist content on social platforms.

2.2 Right-wing radicalism on social media

In recent years, right-wing populists have become increasingly powerful in national and international politics. This can, inter alia, be seen in the latest European Parliament elections in May 2019. EU-critical and nationalist parties achieved 68 of 751 seats in the European Parliament [23].

Another example for the ongoing rise of right-wing populism is the rise of the Alternative für Deutschland (AfD) since its foundation in 2013 [35]. The original aim of the AfD was to abolish the Euro as a shared currency but since the refugee crisis in 2014, their language has drifted further towards the extreme political right. In Germany's federal elections in 2017, AfD achieved 12.6 percent of the vote, up from 4.7 percent four years earlier [8]. It should be emphasized in this context that AfD's popularity is in no small part due to their online presence [35].

In Germany, right-wing populism and radicalism are political ideologies with the following characteristics. Primarily right-wing radicals try to create homogeneity within national borders, and further, to exclude people or groups outside of these borders [33]. Salzborn [[33], p.21-23] also mentions further important characteristics such as a *völkisch* ideology of ethno-nationalism, racism, anti-authoritarianism (refusal to accept any ruling elite), homogeneity-orientation (focus on a collective to which every individual has to subordinate him- or herself), sexism, anti-semitism, anti-Americanism, historical revisionism, militarism and anti-rationalism. Researchers do not agree on a universal definition of right-wing extremism. It is also difficult to delimit the concepts of right-wing populism and right-wing extremism because boundaries between these terms are unclear. This is due to the fact that some characteristics can be attributed to right-wing extremism as well as populism.

In the course of technological progress, the Internet enables a special form of participation for right-wing extremists and populists. To understand why social media platforms such as YouTube may offer easy access to an audience for them, it is necessary to understand the principle of gatekeeping in the context of news dissemination. Since the early years of the 20th century, journalists have seen themselves as professionals with special expertise and a sense of responsibility towards society. Their role in filtering

information became known as the gatekeeper model [17]. In established, traditional media such as newspapers, published content has to overcome this hurdle to be shown to a wider audience [28]. In the case of social media, however, the hurdle of gatekeeping is removed because content is not selected by journalists, but rather by algorithms that are based on user behavior. To be more precise, instead of being reviewed by journalists who weigh ethical concerns for each issue, it is the users themselves who decide if content is worth sharing [19]. This shift has helped to spread and normalize extremist engagement because content published on social media is not monitored with an eye to objective ethical and political concerns [13]. Recent research has shown the influence of the interrelatedness between counter-messages and extremist content based on the YouTube recommendations and videos [34]. Two counter-message campaigns (#WhatIS and ExitUSA) were examined using network analysis. The results showed that there is a certain probability that users will encounter extremist messages because they are related to counter-messages. Medina Serrano [35] also explain how Germany's right-wing populist party AfD uses social media platforms, revealing their knowledge and leveraging of these circumstances. In this case, the party utilizes the fact that most supporters of populism tend to distrust traditional media platforms. Their study focused on alternative media such as social media platforms. Another important aspect is that hate can spread faster in social media, which leads to AfD using a more negative and aggressive rhetoric in their social media channels. The third strategy the authors mentioned is the use of social bots which try to manipulate trends in the interests of AfD.

Since overall there is little research on right-wing radical or populist filter bubbles on YouTube, we question the impact of the YouTube algorithm on the selection of political videos (right-wing populist and neutral) and their recommendations. Therefore, our research questions are:

RQ1: Do users receive recommendations for right-wing populist / neutral videos after watching a right-wing populist / neutral video?

RQ2: Do the recommendations following the consumption of right-wing populist videos also predominantly link to right-wing populist videos?

RQ3: What is the homogeneity of recommendations in the network of neutral and right-wing populist videos?

3 METHODS

3.1 Dataset

To investigate the recommendation behavior of YouTube regarding German political content, we concentrated on ten videos each with neutral political content and right-wing populist content, which used as the set of initial videos. The initial right-wing populist videos were selected through detailed research, and included videos from channels that were identified as belonging to the extreme right-wing spectrum. To determine which video creators spread extreme-right content we searched for reliable indicators. One of the initial videos, for instance, was created by the "Identitäre Bewegung Deutschland" (IBD), which the German Federal Office for the Protection of the Constitution describes as a political movement that infringes on human dignity and the concept of democracy [7].

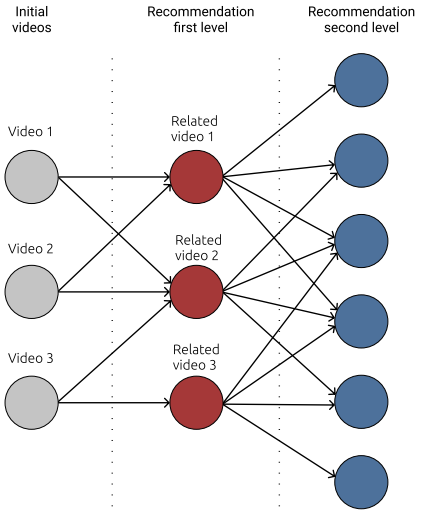

Figure 1: Simplified example of video recommendation connections.

We also chose a video made by the "Junge Freiheit" (JF). Founded as a student newspaper in 1986, the Office for the Protection of the Constitution of North Rhine-Westphalia has monitored the JF since 1994 and describes it as a "hinge between the spectrum of democracy and right-wing extremism" [[5], p.9-11 & p.59]. As another example of a right-wing video creator, we chose the "Deutsche Stimme", the party magazine of the "Nationaldemokratische Partei Deutschland", which spreads right-wing radical and neo-Nazi comments, according to the Office for the Protection of the Constitution of Baden-Wurttemberg [15].

For the initial ten videos in the neutral dataset, only videos from public broadcasters were chosen. These are characterized by a degree of independence from governments and financial interests, the professional requirement for objectivity in journalism, and equal representation of different viewpoints [2].

On July 3, 2019, we performed an automatic data collection via the YouTube API to get information about the initial videos and their first- and second level recommendations. In particular, we used the "*relatedToVideoId*" parameter, which returns a list of recommendations that are related to the original video.[1] Figure 1 shows the relationships between the videos. To avoid any misinterpretation, it is necessary to clarify that not all of the videos in the right-wing populist dataset are right-wing populist videos. Instead, the dataset consists of the ten initial videos, all of which are right-wing populist, and their first- and second level recommendations (which themselves are not necessarily right-wing populist).

Data collection via the YouTube API ensured that no personal, local, geographic or other variable affected the data collection of the

[1] https://developers.google.com/youtube/v3/docs/search/list

Table 1: Indicators of the total dataset (including duplicates).

Dataset	Total views	Total likes	Total dislikes	Total comments	Total videos
Right-wing populist	425,242,090	5,464,155	863,008	1,171,123	1,118
Neutral	956,529,409	9,089,595	1,006,244	1,385,725	1,076

recommendations. These recommendations lead back to YouTube's algorithm, which suggests videos to the user based on their personal viewing behavior. Based on ten politically neutral videos, 100 recommendations at the first level and 966 recommendations at the second level were collected. The ten populist right-wing videos lead to 100 videos at the first and 1,008 recommendations at the second level. The following shows the total popularity indicators of the videos from the initial sets and the recommendations.

3.2 Annotation of the videos

Unfortunately, the data collected via the API does not reveal whether recommended videos are right-wing populist, politically neutral or something else, so we proceeded to encode the data manually. Following a fixed coding scheme, we annotated the YouTube videos according to three different classes. The data were labelled with one of three classes: "right-wing populist", "politically neutral" and "others". Table 2 shows the classes of videos with their descriptions and characteristics.

Videos were categorized as *right-wing populist* when they contained populist or radical content and implied hostility toward women, foreigners or the LGBTQ community. The class *neutral* was assigned to those videos that reported on politics objectively. This is characterized by showcasing several points of view, and various opinions being presented alongside one another, such as in documentaries and journalistic reports. Videos were categorized as *others* if they did not have a clear political reference and, for example, were more for entertainment purposes, such as gameplay videos and lifestyle blogs. Left-wing videos were also included in the category *others*, due to the fact that only a limited number of these videos were present in the dataset and they were therefore not relevant for further analysis or to our research questions.

According to YouTube's guidelines[2], channels or individual videos are banned if they contain an incitement to hatred or violence against specific individuals or groups. The guidelines explicitly mention characteristics such as ethnic origin, gender identity, nationality, religion or gender. In the period between retrieving and annotating the dataset, some videos had been removed by YouTube, which means it was not possible to watch them anymore. Most of the deletions were because of hate speech. In cases where channel owners decided to restrict videos to a private audience, they were categorized using the video's metadata provided by the API (caption and video description). During the annotation, two videos were removed because they were not in German. Furthermore, there were inconsistencies between the first level and second level of the right-wing populist dataset. For this reason, the affected videos (seven in the first level and 70 in the second level) were removed from the dataset. To quantify the quality of the annotations and coding scheme, we performed a reliability test with a randomly generated dataset at each level with 26 videos, which was then coded by another annotator. The inter-coder reliability was measured using Cohens kappa [10]. For both datasets it showed moderate level of agreement (0.505 for the right-wing populist dataset, 0.454 for the neutral dataset).

[2]YouTube guidelines: https://www.youtube.com/about/policies/#community-guidelines

3.3 Building a recommendation network on YouTube

To build a network structure from the collected YouTube data, it was necessary to first transform it into a video recommendation network. We created a separate network for each dataset. It must be noted that the number of nodes does not correspond to the absolute number of videos in the two samples, since duplicate video IDs are combined in the course of data processing. The network properties for each of the two networks are illustrated in Table 3

3.3.1 Recommendation Behavior using random walk. To simulate the behavior of a user in the network and how he or she transitions from the main videos to other recommended videos, a simulation was performed using a random walk. The nature of video recommendations in the political context of right-wing populist videos can be better understood by using such a simulation. This mathematical method uses random steps to retrace the behavior of a user who continues to watch videos by following recommendations. The term random walk refers to a process in which a random sequence of nodes is generated as a result of following edges in the graph at random. In an unweighted graph, the edge to be followed is selected uniformly from all outgoing links at each step, while in a weighted graph the probabilities of the outgoing links are proportional to the link weights. Random walks are used in studying many aspects of online social networks, such as community detection [22, 31], fake account detection [9, 18] and the study of recommendation systems [16, 24]. Recent work has also been done in applying random walks on YouTube to analyze its recommendation engine [36]. This study's findings, based on 14,509 popular English-language channels, indicate that YouTube tends to recommend longer and more popular content to users, regardless of the rating, relevance, date or number of views of the video.

Our simulation was run on the annotated dataset that coded the videos for their political content. In this process, we assume that the user starts with the initial right-wing populist or neutral videos and his or her viewing continues by visiting one video from among the recommendations, until the last recommended video ends. The random walk works as follows:

- 1. Start with a video from the list of ten initial videos.
- 2. Randomly select a related video (one out of ten). At each step, the walker only follows direct edges and chooses an edge from the current node's neighbors uniformly at random.

Table 2: Coding scheme for YouTube videos.

Class	Explanation
Right-wing populist	Right-wing populist or radical content
	Polemical and polarizing speech
	Lack of evidence of the assumptions made or unrealistic assumptions (e.g. protagonist shows xenophobia and misogyny, hostility towards the LGBTQ+ scene and political opponents)
	Rejection/agitation against a certain group, political opponents (esp. the left) (e.g. right-wing extremists, Nazi symbols in the video)
	One-sided presentation of conspiracy theories as facts; denial of anthropogenic climate change
Neutral	Politically unbiased reporting such as documentaries or reports on political issues
	No clear right- or left-wing bias in reporting
	Highlighting different aspects of the content
	Editorially prepared contents
	Protagonist takes an unbiased position, illuminates several sides
Others	Content without political reference such as blogs, vlogs, beauty, lifestyle and gameplay
	Left-wing content

Table 3: Network parameters.

Network parameter	Initial neutral political videos	Initial right-wing videos
Nodes	852	811
Edges	1064	1029
Avg. degree	1.25	1.27
Diameter	7	5
Density	0.0015	0.0016
Max in-degree	10	9
Max out-degree	10	20

- 3. After selecting the related video from the initial video, continue by selecting the next related video randomly (one out of ten).
- 4. Store the sequence of nodes and their attributes for each run.
- 5. Repeat this process 5000 times for each initial video, until all initial videos have been passed.

Since the various recommendations at the first and second levels generate new interconnections between each other, this process is different from only looking at the first and second levels. We decided to set the maximum number of walk steps until the algorithm determines to a value of 2.

3.3.2 Measuring video content homogeneity. In order to measure the homogeneity of the network and to determine to what extent videos in the same political class are recommended for further videos, we used the E-I index introduced by Krackhardt and Stern in 1988 [20]. Thus, it was possible to determine the direct connections from the individual nodes to their recommendations based on the class. The formula of the global E-I Index is defined as:

$$EI = \frac{E - I}{E + I}$$

where E is the number of external links, whereas I is the number of internal links to a node. The E-I index takes values between -1 and 1, where a value of 1 indicates a heterogeneous network and a value of -1 indicates a homogeneous network. An E-I index of 0 indicates a network in which there are as many connections between the groups as there are within the groups. However, when groups are different sizes, an E-I index different from 0 is to be expected even in the absence of homophily, when nodes are connected randomly to other nodes ignoring group membership. This is because the probability that a recommended video is from the same group is inherently higher the larger the group is, even when the video is chosen purely at random.

Therefore, a permutation test was used that measures whether the observed E-I index differs significantly from the one expected under the null hypothesis that the nodes are connected randomly (and that the recommendation system therefore does not favor videos of one political persuasion over another). The sampling distribution was obtained by rewiring the edges while preserving the graph's degree distribution.

Recent studies already applied the index to discussion networks on YouTube to compute the homogeneity analysis of user-generated messages [32]. Bruns [6] suggests that the investigation of filter bubbles and echo chambers should focus on network analysis techniques such as the usage of the E-I index in order to analyze the communication structure of the information between the individual users. Especially in the online, connected world we live in today, where algorithms are not transparently communicated by companies, it is important to investigate and compare the behavior of this

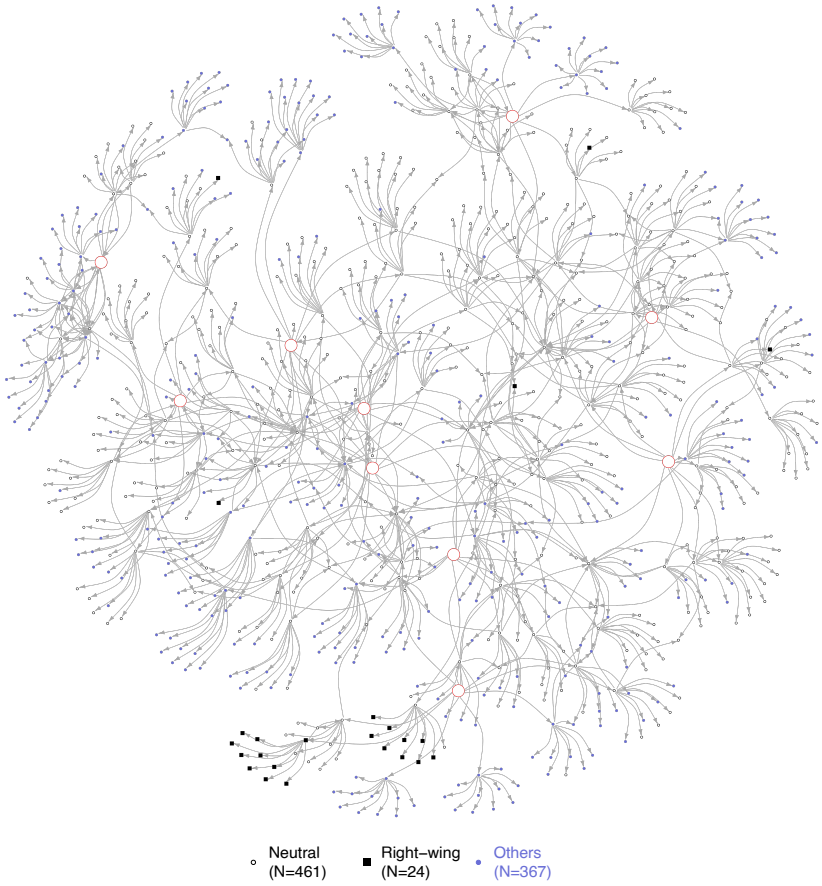

Figure 2: Recommendation network of the initial neutral dataset. The ten initial videos are identified by the size of the vertex and the red border.

recommendation system with respect to the political context and especially to right-wing populist content in the network.

4 RESULTS

The following Figure 2 and Figure 3 demonstrate the recommendation networks of videos which had ten populist right-wing and ten neutral political videos as a starting point, and which include their recommendations at the first and second level. The nodes of the network are the individual YouTube videos. They are colored according to their content: white for neutral, blue for "others" and black for populist right-wing content.

Regarding RQ1, more than half (53.76 percent) of recommendations received after watching one of the initial right-wing populist videos were other right-wing populist videos. However, more than a third (36.56 percent) of recommended videos were politically neutral. Less than one in ten (9.68 percent) recommendations belonged to the category *others*. In contrast, when the initial video is politically neutral, only two percent of recommendations are right-wing populist videos, while 73 percent are neutral themselves and 25 percent belong to the *others* category.

Addressing RQ2, the results show the recommendations of videos differ between the first and second recommendation levels. In the

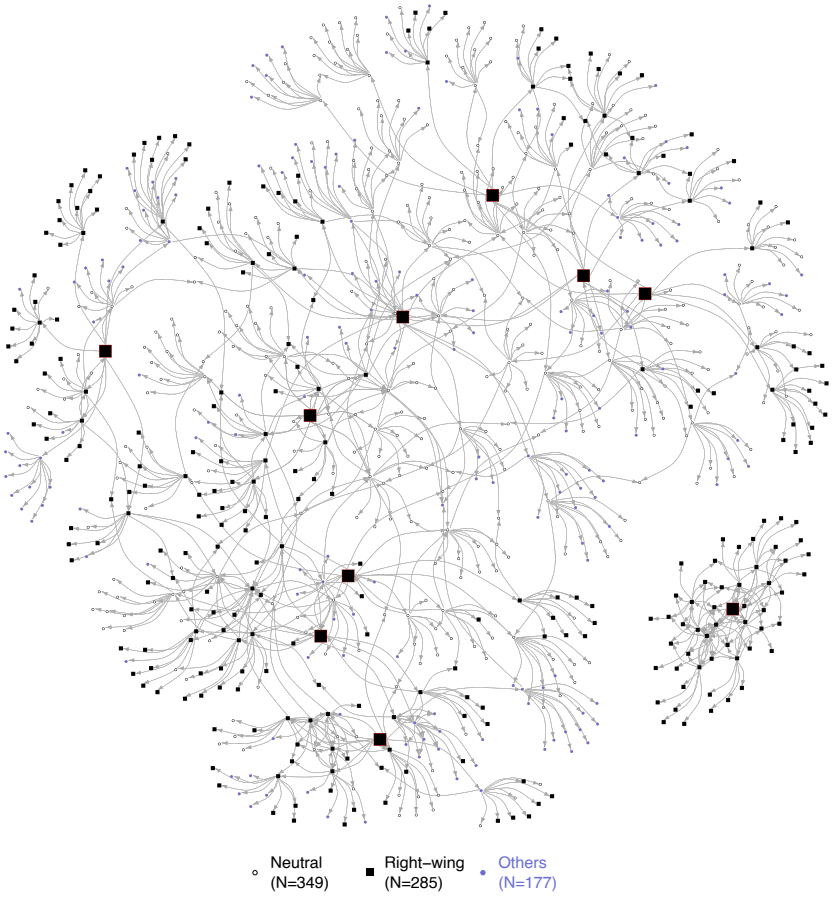

Figure 3: Recommendation network of the initial right-wing dataset. The ten initial videos are identified by the size of the vertex and the red border.

right-wing dataset, only 37.71 percent of recommendations at the second level are right-wing videos, while the share of neutral (42.2 percent) and *others* (20.09 percent) videos increases. Comparing the second and first recommendation levels, the frequency of right-wing populist videos decreases. Thus, fewer right-wing populist videos are found at a higher recommendation level. In the neutral dataset 55.5 percent of recommendations are neutral and 41.91 percent are *others*, whereas only 2.59 percent of recommendations are right-wing populist videos.

In addition to the real networks and their actual class probabilities in Figure 3, it is relevant to determine whether and at what frequency a user comes into contact with these different political videos. We assume that the user has previously visited one of the initial politically neutral or right-wing populist videos. Looking at the two networks individually, a noticeable trend can be seen.

In the neutral network, walkers mainly encounter neutral (65.73 percent) and other (32.53 percent) videos, while right-wing populist (1.74 percent) videos are not frequently visited. Comparing the walker's steps with the videos the user has visited, these right-wing populist videos are mainly watched immediately after the initial video or as a following second level recommendation. These results are in line with the measured probabilities.

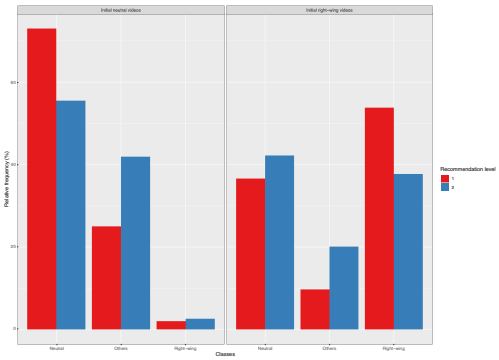

Figure 4: Relative frequency of videos on different recommendation levels on YouTube.

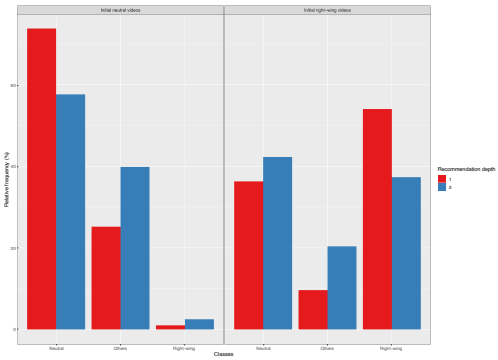

Figure 5: Relative frequency of the random walk simulation.

When the initial videos are right-wing populist, however, it becomes clear that subsequently, right-wing populist videos (45.7 percent) are most frequently visited by the walkers, followed by neutral (39.3 percent) and other (15 percent) videos. When looking at the individual steps, it is seen that immediately after the first video, the probability of watching a right-wing populist video is highest. In the second recommendation level, however, the neutral videos are visited most often. Right-wing populist videos are encountered less frequently. The results of the random walk for both networks are in line with the actual probabilities in Figure 4 and can therefore realistically represent the behavior of the user in the network. Figure 5 shows a summary of the results of the random walk.

Regarding the homogeneity of the network, we found that both networks are indicated with a negative E-I index, which means that the network has more like-minded than cross-cutting connections. The global E-I index for initial right-wing populist videos was -0.337, which is calculated from the 341 external and 688 internal links. The results of the network with neutral political videos show an even more negative E-I index of -0.508. It is calculated from the 262 external and 802 internal connections.

The results of the permutation tests show that a negative E-I index is to be expected given the differences in the sizes of the three groups. However, the observed E-I index is significantly closer to the theoretical extreme of -1 than what would be expected under the null hypothesis.

The right-wing populism network shows that a separate isolated community of right-wing populist videos has formed within the network, which is not connected to the other videos. This is alarming because it could lead to a spiral of like-minded video consumption if users follow these recommendations. The YouTube recommendation algorithm partially paves the way for staying on the politically extreme path, especially if the user has had the impulse to visit something politically extreme from the beginning.

5 DISCUSSION

5.1 Filter bubbles on YouTube

The goal of this research was to examine whether YouTube's recommendation algorithm contributes to creating a "filter bubble" regarding political ideology, especially German right-wing populist content.

The collected data allows us to answer the research questions posed earlier. Regarding RQ1, viewers of right-wing populist videos are indeed much more likely to be shown recommendation for more right-wing populist content (53.76 percent) than recommendations for politically neutral videos. However, more than a third (36.56 percent) of recommendations were still politically neutral videos. Viewers of politically neutral videos, on the other hand, are very unlikely (2 percent) to be recommended right-wing populist content. RQ2 asked whether the recommendations next to these recommended videos (i.e., depth 2) are also predominantly right-wing populist videos. At this level, viewers who initially start with a right-wing populist video are actually more likely to be recommended a neutral video than a third right-wing populist one. The analysis for RQ3 showed that the network of recommendations is significantly more homogeneous than what would be expected if the recommendations were random, in other words, if they were blind to the political viewpoints that are espoused in the videos. In summary, the recommendations showed a tendency in favor of showing users more of the political content that they had already watched, but the random walk showed that it would be inaccurate to say that they were led down rabbit holes that are impossible to escape.

From YouTube's point of view, these results probably indicate that the recommendation system works as intended. It should be emphasized that our results do not suggest that the recommendation algorithm explicitly takes the political views in a video into account. The observed phenomena can simply be explained by the fact that these algorithms are designed to maximize user engagement and user satisfaction. Just as it learns that users who watch gaming videos are more likely to continue watching if they are recommended more gaming videos as opposed to, for example, music

Table 4: Results of the E-I index based on the permutation test with 5000 iterations.

Dataset	E-I index	Min	Avg. E-I	Max	SD	p
Right-wing populist	-0.337	0.133	0.237	0.328	0.028	<0.001
Neutral	-0.508	-0.145	-0.057	0.021	0.025	<0.001

videos, the system has probably decided, based on the available data, that people who consume right-wing populist content are more likely to continue watching if they are recommended more populist content.

However, there is a fundamental difference between political content and many of the other types of content that can be found on YouTube. People form political opinions and voting intentions based on the political news they consume. In the past, journalists acted as gatekeepers for the information presented to the public, for better or worse. This is no longer the case. Truthfulness, independence and accountability are core values in traditional journalism ethics. By contrast, the unhindered spread (e.g., on social media) of misinformation from dubious sources, who may have political intentions of their own, has been theorized to be a threat to democracy.

Changing the status quo raises countless questions about algorithm design, transparency, and regulation. Pro-populist "bias" we have demonstrated in the recommendation algorithm has been learned from the data, just like the anti-populist "bias" for users who prefer to consume politically neutral videos. It might be possible for developers to encode a distinction between hobbies and leisure activities, and political content, thereby stopping the system from applying the same heuristics to these two distinct realms. This may yield results that are less optimal in terms of predicted engagement and user satisfaction, but perhaps better in terms of the journalistic quality of the news selection or some other, ideally quantifiable measure. Certainly, such a decision would need to be conscious and be made by software designers or lawmakers. If social media platforms begin implementing systems that allow the spread of some political views but contain or stop the spread of others, how do we draw the line between desirable views and undesirable ones? Who ensures that such a system is not abused in authoritarian states to stop the views of the political opposition from spreading? One can imagine that over-regulation could have unintended side effects.

As a result of our data collection method, the YouTube recommendation algorithm only had access to the previously watched video to calculate its recommendations. This allowed us to study the effect of the previously watched video in isolation. In a real setting, recommendations are likely to be influenced by a wide range of other factors, such as a user's watch history, search history, and other data. Characteristics of videos are also likely to play a role, for example the total number of views and likes.

5.2 Limitations

The present work is subject to certain limitations. The first limitation of this study is that the data was collected under the controlled conditions of the YouTube API. This has the advantage that the recommendations are not tailored based on personal information such as videos viewed in the past in a browser, which would endanger the generalizability of the results. However, the data collection does not reflect an "average" user who has a stored history of keyword searches and watched videos. Under real conditions, users' previous viewing behavior and other factors will also influence the recommendations made by the algorithm. For this reason, a long-term study involving several users with the same search history and criteria, in which the algorithm's filtering and recommendation levels are investigated, would be relevant to compare the results with each other.

Secondly, we only investigated two recommendation levels with ten initial videos each. A larger dataset would help to analyze an even more interconnected network structure and to gain further insights into deeper recommendation levels.

Another limitation of the study is related to the annotation. By the time the data annotation was complete, some of the videos collected from YouTube had been deleted due to the platforms guidelines on violations (e.g. due to hate speech). The fact that some videos were blocked by YouTube nevertheless implies something important: YouTube's enforcement of hate speech policies against channel operators indicates that the threat posed by right-wing populist and radical contributions to the platform is recognized by the company. In future work, researchers will need to endeavor to keep the time lag between data collection and subsequent annotation as short as possible, to ensure that the data is not "lost" for further analysis.

As a further limitation of the work it should be mentioned that only one additional annotator was used for the computation of the inter-coder reliability. A higher number of annotators would therefore increase the quality of the results and allow for a more detailed examination.

6 CONCLUSION & FURTHER WORK

This study investigated political content on YouTube and how it is linked in a recommendation network to evaluate the filter bubble hypothesis. We labeled the videos as belonging to one of three categories and examined two levels of recommendations in more detail. We found that when a user begins by watching a right-wing video, right-wing populist videos are more frequently suggested at the first recommendation level than at the second level. Politically neutral videos predominantly link to other neutral or "others" videos. Furthermore, we used the E-I Index to investigate the homogeneity in the network. Results indicate that the recommendation network shows a highly homogeneous behavior of the examined classes.

Our research examined the YouTube recommendation network and how right-wing populist and neutral political videos are linked. Further research could investigate user reactions in the form of user-generated comments to these videos to analyze the political opinions of users and thus the opinion climate. This analysis would reveal the extent to which comments indicate a focus on the active

behavior of users (in contrast to passive consumer behavior). Another issue that was not addressed in this study was the intensity of the content (e.g., how extreme are the political videos in a video?). Based on this information, future studies could determine whether users are confronted with increasingly radical political content, which would be a different manifestation of the filter bubble to the one researched in this study.

ACKNOWLEDGMENTS

This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia (Grant Number: 005-1709-0004), Junior Research Group "Digital Citizenship in Network Technologies" (Project Number: 1706dgn009).

REFERENCES

[1] Adiya Abisheva, David Garcia, and Frank Schweitzer. 2016. When the filter bubble bursts: collective evaluation dynamics in online communities. In *Proceedings of the 8th ACM Conference on Web Science - WebSci '16*, ACM Press, Hannover, Germany, 307–308. DOI:https://doi.org/10.1145/2908131.2908180

[2] ARD. 2018. Öffentlich-rechtlicher Rundfunk. Retrieved from http://www.ard.de/home/die-ard/fakten/Oeffentlich_rechtlicher_Rundfunk/458368/index.html

[3] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (June 2015), 1130. DOI:https://doi.org/10.1126/science.aaa1160

[4] Engin Bozdag and Jeroen van den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics Inf Technol* 17, 4 (December 2015), 249–265. DOI:https://doi.org/10.1007/s10676-015-9380-y

[5] Stephan Braun and Ute Vogt. 2007. *Die Wochenzeitung „Junge Freiheit": Kritische Analysen zu Programmatik, Inhalten, Autoren und Kunden*. DOI:https://doi.org/10.1007/978-3-531-90559-4

[6] Axel Bruns. 2019. It's not the technology, stupid: How the 'Echo Chamber' and 'Filter Bubble' metaphors have failed us. Retrieved from https://eprints.qut.edu.au/131675/

[7] Bundesministers des Innern, für Bau und Heimat. 2018. *Verfassungsschutzbericht 2018*. Retrieved from https://www.verfassungsschutz.de/embed/vsbericht-2018.pdf

[8] Bundeswahlleiter. 2017. Wahl zum 19. Deutschen Bundestag am 24. September 2017. Heft 3. Endgültige Ergebnisse nach Wahlkreisen. Bundeswahlleiter Wiesbaden.

[9] Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. 2012. Aiding the Detection of Fake Accounts in Large Scale Social Online Services. In *Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, USENIX, San Jose, CA, 197–210. Retrieved from https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/cao

[10] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[11] Flavio Figueiredo, Fabrício Benevenuto, and Jussara M. Almeida. 2011. The tube over time: characterizing popularity growth of youtube videos. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, ACM Press, Hong Kong, China, 745. DOI:https://doi.org/10.1145/1935826.1935925

[12] Paolo Gerbaudo. 2018. Social media and populism: an elective affinity? *Media, Culture & Society* 40, 5 (July 2018), 745–753. DOI:https://doi.org/10.1177/0163443718772192

[13] Nitin Govil and Anirban Kapil Baishya. 2018. The Bully in the Pulpit: Autocracy, Digital Social Media, and Right-wing Populist Technoculture. *Communication, Culture and Critique* 11, 1 (2018). DOI:https://doi.org/10.1093/ccc/tcx001

[14] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. 2018. Burst of the Filter Bubble?: Effects of personalization on the diversity of *Google News*. *Digital Journalism* 6, 3 (March 2018), 330–343. DOI:https://doi.org/10.1080/21670811.2017.1338145

[15] Innenministerium Baden-Württemberg. 2015. DIE NPD-PARTEIZEITUNG „DEUTSCHE STIMME". In *Verfassungsschutz 2014 Baden-Württemberg*. 198–200. Retrieved from https://im.baden-wuerttemberg.de/fileadmin/redaktion/m-im/intern/dateien/pdf/Verfassungsschutzbericht2014_web_Juli2015.pdf

[16] Mohsen Jamali and Martin Ester. 2009. *TrustWalker*: a random walk model for combining trust-based and item-based recommendation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, ACM Press, Paris, France, 397. DOI:https://doi.org/10.1145/1557019.1557067

[17] Morris Janowitz. 1975. Professional models in journalism: The gatekeeper and the advocate. *Journalism quarterly* 52, 4 (1975), 618–626.

[18] Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. 2017. Random Walk Based Fake Account Detection in Online Social Networks. In*2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, IEEE, Denver, CO, USA, 273–284. DOI:https://doi.org/10.1109/DSN.2017.55

[19] Till Keyling. 2017. *Kollektives Gatekeeping: Die Herstellung von Publizität in Social Media*. Springer VS, Wiesbaden.

[20] David Krackhardt and Robert N. Stern. 1988. Informal Networks and Organizational Crises: An Experimental Simulation. *Social Psychology Quarterly* 51, 2 (1988), 123–140. DOI:https://doi.org/10.2307/2786835

[21] Mark Ledwich and Anna Zaitsev. 2019. *Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization*.

[22] Weimin Li, Jun Xie, Mingjun Xin, and Jun Mo. 2018. An overlapping network community partition algorithm based on semi-supervised matrix factorization and random walk. *Expert Systems with Applications* 91, (January 2018), 277–285. DOI:https://doi.org/10.1016/j.eswa.2017.09.007

[23] BW LPB. 2019. *Europa hat gewählt, Landeszent-rale für politische Bildung Baden-Württemberg*. Retrieved from https://www.europawahl-bw.de/wahlergebnis_europawahl2019.html

[24] Yijun Mo, Bixi Li, Bang Wang, Laurence T. Yang, and Minghua Xu. 2018. Event recommendation in social networks based on reverse random walk and participant scale control. *Future Generation Computer Systems* 79, (February 2018), 383–395. DOI:https://doi.org/10.1016/j.future.2017.02.045

[25] German Neubaum and Nicole C. Krämer. 2017. Monitoring the Opinion of the Crowd: Psychological Mechanisms Underlying Public Opinion Perceptions on Social Media. *Media Psychology* 20, 3 (July 2017), 502–531. DOI:https://doi.org/10.1080/15213269.2016.1211539

[26] Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. 2013. The Extreme Right Filter Bubble. *arXiv:1308.6149 [physics]* (August 2013). Retrieved January 16, 2020 from http://arxiv.org/abs/1308.6149

[27] Raphael Ottoni, Evandro Cunha, Gabriel Magno, Pedro Bernardina, Wagner Meira Jr., and Virgílio Almeida. 2018. Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination. In *Proceedings of the 10th ACM Conference on Web Science - WebSci '18*, ACM Press, Amsterdam, Netherlands, 323–332. DOI:https://doi.org/10.1145/3201064.3201081

[28] Ute Pannen. 2010. Social Media: Eine neue Architektur politischer Kommunikation. *Forschungsjournal Soziale Bewegungen* 23, 3 (January 2010). DOI:https://doi.org/10.1515/fjsb-2010-0308

[29] Eli Pariser. 2012. *The filter bubble: what the Internet is hiding from you*. Penguin Books, London.

[30] A Perrin and M Anderson. 2019. Share of US adults using social media, including Facebook, is mostly unchanged since 2018. Pew Research Center.

[31] Pascal Pons and Matthieu Latapy. 2005. Computing Communities in Large Networks Using Random Walks. In *Computer and Information Sciences - ISCIS 2005*, Springer Berlin Heidelberg, Berlin, Heidelberg, 284–293.

[32] D. Röchert, G. Neubaum, B. Ross, F. Brachten, and S. Stieglitz. 2020. Opinion-based Homogeneity on YouTube: Combining Sentiment and Social Network Analysis. *Computational Communication Research* 2, 1 (2020), 81–108. DOI:https://doi.org/10.5117/CCR2020.1.004.ROCH

[33] Samuel Salzborn. 2015. *Rechtsextremismus: Erscheinungsformen und Erklärungsansätze* (2., aktualisierte und erweiterte Auflage ed.). Nomos, Baden-Baden.

[34] Josephine B Schmitt, Diana Rieger, Olivia Rutkowski, and Julian Ernst. 2018. Counter-messages as Prevention or Promotion of Extremism?! The Potential Role of YouTube. *Journal of Communication* 68, 4 (August 2018), 780–808. DOI:https://doi.org/10.1093/joc/jqy029

[35] Juan Carlos Medina Serrano, Morteza Shahrezaye, Orestis Papakyriakopoulos, and Simon Hegelich. 2019. The Rise of Germany's AfD. In *Proceedings of the 10th International Conference on Social Media and Society - SMSociety '19*, ACM Press, New York, New York, USA, 214–223. DOI:https://doi.org/10.1145/3328529.3328562

[36] A Smith, S Toor, and P Van Kessel. 2018. Many Turn to YouTube for Children's Content, News, How-To Lessons. *Pew Research Centre* (2018).

[37] Cass R. Sunstein. 2017.*#Republic: divided democracy in the age of social media*. Princeton University Press, Princeton; Oxford.

[38] Zhe Zhao, Ed Chi, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, and Xinyang Yi. 2019. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM Conference on Recommender Systems - RecSys '19*, ACM Press, Copenhagen, Denmark, 43–51. DOI:https://doi.org/10.1145/3298689.3346997

[39] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. 2010. The impact of YouTube recommendation system on video views. In *Proceedings of the 10th annual conference on Internet measurement - IMC '10*, ACM Press, Melbourne, Australia, 404. DOI:https://doi.org/10.1145/1879141.1879193

[40] Frederik Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balázs Bodó, Claes H De Vreese, and Natali Helberger. 2016. Should we worry about filter bubbles? *Internet Policy Review. Journal on Internet Regulation* 5, 1 (2016).

Research Paper 6: "Two sides of the same leader: an agent-based model to analyze the effect of ambivalent opinion leaders in social networks"

Type	Journal
Rights and permission	Reproduced with permission from Springer Nature Open access
Authors	**Röchert, Daniel**; Cargnino, Manuel; Neubaum, German
Year	2022
Outlet	Journal of computational social science
Publisher	Springer Nature
Permalink/DOI	https://doi.org/10.1007/s42001-022-00161-z
Full citation	Röchert, D., Cargnino, M., & Neubaum, G. (2022). Two sides of the same leader: an agent-based model to analyze the effect of ambivalent opinion leaders in social networks. *Journal of computational social science*, 1-47.

Journal of Computational Social Science
https://doi.org/10.1007/s42001-022-00161-z

RESEARCH ARTICLE

Two sides of the same leader: an agent-based model to analyze the effect of ambivalent opinion leaders in social networks

Daniel Röchert[1] · **Manuel Cargnino**[1] · **German Neubaum**[1]

Received: 1 July 2021 / Accepted: 16 February 2022

Abstract
Opinion leaders (OLs) are becoming increasingly relevant on social networking sites as their visibility can help to shape their followers' attitudes toward a variety of issues. While earlier research provided initial evidence on the effect of OLs using agent-based modeling, it remains unclear how OLs affect their network environment and, therefore, the opinion climate when: (a) they publicly hold ambivalent attitudes, and (b) they not only express support for their own stance but also discredit or 'debunk' the opposing side. This paper presents an agent-based model that determines the influence of OLs in social networks in relation to ambivalence and discreditation. The model draws on theoretical foundations of OLs as well as attitudinal ambivalence and was implemented using two network topologies. Results indicate that OLs have significant influence on the opinion climate and that an unequal number of OLs of different opinion camps lead to an imbalance in the opinion climate only in certain situations. Furthermore, OLs can dominate the opinion climate and turn their stance into a majority opinion more effectively when discrediting the opposing side. Ambivalent OLs, on the other hand, can contribute to greater balance in the opinion climate. These findings provide a more nuanced analysis of OLs in social networks by pointing to potential amplifications as well as boundaries of their influence. Implications are discussed with a focus on human and artificial key actors in online networks and their efficacy therein.

Keywords Opinion leader · Agent-based modeling · Ambivalence · Simulation · Network analysis

✉ Daniel Röchert
daniel.roechert@uni-due.de

[1] University of Duisburg-Essen, Duisburg, Germany

Published online: 26 April 2022

Introduction

The rise of digital communication platforms such as social networking sites has been posing new challenges in terms of understanding processes of opinion formation. The connectedness between users and the exchange of political information provides vast possibilities of mutual social influence. On social networking sites, opinion leaders (OLs) play a key role, as they can have a disproportionate influence on the opinions in their environments and thus the prevailing opinion climate. OLs are characterized by their high connectedness in networks, which enables them to steer the diffusion of information in a certain direction [1, 2]. However, on social networking sites, the influence of OLs is embedded in complex communication settings. For instance, OLs and other individuals do not always hold opinions that clearly favor one side. Instead, they can hold and express ambivalent opinions that equally favor both sides [3]. When it comes to OLs, public service broadcasters and non-partisan journalists may be considered ambivalent as they are bound to balanced reporting. Furthermore, public opinion or a network as a whole can be ambivalent when opinions on a political question vary between citizens ("network ambivalence" [4]). However, previous research has conceptualized opinion dynamics in social networks as the likelihood of expressing the valence of one's opinion (i.e., to be in favor *or* against a political decision) without accounting for compelling social psychological evidence that indicates that individuals' opinions are often not purely supporting or opposing an issue but can be ambivalent [5]–[7]. On social networking sites, the factual argumentation of statements is not always in the foreground, but instead, vulgarities and "dirty tricks" oftentimes characterize communication [8]. Public advocates of one viewpoint do not only talk about their stance, but also make references to "the other side," be it in the form of counterarguing, providing substantial arguments, or even discrediting the credibility (e.g., the expertise and trustworthiness) of opponents and their views [9].

These two observations (i.e., the ambivalent and discrediting expressions of members of a network) have implications for how opinion climates evolve on social networking sites. Despite growing knowledge gained through agent-based modeling on the mechanisms that drive changes in opinion climates [10, 11], the observation that opinions (even those propagated by OLs) can be ambivalent or that key agents can discredit "the other side" have not been implemented in agent-based modeling to date. Taking these into account appears to be of pivotal relevance when it comes to explaining complex dynamics in online discussion networks. OLs may not necessarily advocate a clear and one-sided stance (e.g., spreaders of partisan media content), but instead convey equilibrated, i.e., ambivalent stances (e.g., spreaders of mainstream media or public service broadcasters). Including these aspects in the simulation of opinion formation processes allows for the modeling of more complex and realistic processes in large social networks. To our knowledge, there are no agent-based models to date in which attitude ambivalence is applied to OLs and users in networks. Agent-based models are best suitable to address social phenomena by simulating individual behavior and observing its impact on a group/network level [12].

Against this background, the present study is intended to use virtual simulations to: (1) include ambivalent opinions in the modeling of complex social influence processes in social networks; (2) implement ambivalence also in OLs' expression behavior; and (3) take into account not only supportive expressions of viewpoints by OLs, but also discrediting utterances.

With that said, our work first acknowledges that OLs can be present simultaneously for different opinion camps. Still, it is yet to be understood whether variations in the number of OLs in different factions can make a difference in the ultimate outcome of the opinion climate. Therefore, we ask:

RQ 1: How does the opinion climate respond to a varying ratio of OLs in each group?

Second, we are interested in testing the effects of OLs on their network when they not only express one exclusive stance but also convey ambivalence in their utterances. While related work has shown that a like-minded social network environment can lead to a strengthening of users' opinions [13] and underlines the important role of OLs in the process of diffusion [14], little is known on the impact of ambivalent opinions conveyed by OLs. Therefore, we aim to compare effects of univalent and ambivalent OLs on the network level and ask:

RQ 2: How does the opinion climate respond to OLs who advocate fully in favor of their stance, are fully ambivalent, or partly in favor of the opponents' stance?

Third, we also intend to examine the situation in which an OL not only presents support for a stance but also actively discredits or debunks the opposing side:

RQ3: How does the opinion climate respond to OLs of one side who advocate fully in favor of their stance and discredit the opponents' stance?

Theoretical background

In this section, we outline the theoretical background of OLs and how they have already been implemented in agent-based models. Furthermore, we discuss the theory behind attitudinal ambivalence as well as how related psychological processes can be applied to the model of OLs.

Opinion leaders in social networks

The networking of individuals in social networking sites facilitates not only rapid communication among users, but also exposure to a variety of information and different opinions. OLs can play a key role in this process, as they are not only connected to numerous users in the network, but also have a strong influence on other users' opinions. "Generally, opinion leader is obviously the critical node with a higher centrality in social networks, and it is bound to affect public opinion assimilation, integration, and separation, but his/her influence on opinion evolution is

obviously different from that of ordinary individual […]" [15, p. 3]. Lazarsfeld et al. [16] introduced the notion of OLs while studying US presidential elections. They investigated electoral behavior and proposed a two-step flow model: mass media indirectly influence the general public by first providing information to OLs who, in turn, pass this information on to individuals [16, 17].

Due to the fact that the use of social networking sites has been increasing in recent decades, research has started to study how OLs and their influence manifest themselves in social networking platforms. Studies have shown that OLs play a major and central role in the dissemination of information in online discussions [14], having influence on individuals [18] even if they are politically uninterested [19]. A recent study based on the social media debate on climate change showed that political actors are highly qualified to fulfill the position of OLs, as they have a significant impact on the flow of information, as characterized by the fact that they posted a higher volume of tweets and were also more frequently mentioned by other users [20].

The concept of OLs is often associated with information diffusion, which describes the process of how information is spread within the network. Based on results focusing on OLs and diffusion processes, research showed that the presence of only a few OLs can have an impact on how quickly information is spread in online networks [21, 22]. Simulation studies have identified that OLs with high "sociality" (i.e., the total strength of ties of an entity) are best placed to rapidly disseminate information. However, OLs only influence the diffusion process if the percentage of first-time addressees reaches a critical mass [22]. Previous studies have shown that the proportion of OLs varies and depends on the object of investigation. While Choi [23] identified a proportion of only 4% of OLs in a study on Twitter-based discussion groups in South Korea, previous surveys on offline OLs identified 23%–30% of respondents as OLs [24]. A similar distribution was also used in an agent-based simulation [25, p. 201]. Finally, a detailed examination by Weeks [2] identified 12.5% of users as OLs.

While these findings imply that the number of OLs varies, little is known on the impact that different proportions of OLs might have on the dynamics of opinion formation and resulting opinion climates in online networks.

Influence: conceptual issues

Related work on processes of opinion formation and influence through influential players (i.e., OLs) used different concepts and operationalizations. For instance, Bakshy et al. [26] studied influence among Twitter users based on network diffusion. Accordingly, influence is understood as the degree to which a piece of information is spread through the network. The wider a piece of information is spread (i.e., the larger the size of the "diffusion tree"), the more influential its sender. According to this view, influence is something that can be directly observed through the extent of information diffusion [20], which makes OLs influential players, since information stemming from them is of relatively high reach. The present work addresses influence mainly on the level of opinions or attitudes toward a specific issue (e.g., a controversial policy), i.e., on the level of internal psychological processes. On the

network level, influence is thus characterized by the degree to which a user's opinion is impacted by opinions represented in their environment. Consequently, OLs are influential, since opinions propagated by them have a relatively high impact on neighboring opinions [2]. This concept of influence most closely resembles models of opinion formation (in particular, the two-step flow model, see [13, 14, 22]) and psychological models of social influence (e.g., [27, 28]), which refer to the role of influential single actors (i.e., OLs) and groups (e.g., a local environment within a social network) in processes of opinion formation, respectively.

Agent-based modeling of opinion dynamics

With a focus on the actual communication process, there is already a vast body of research that employs agent-based modeling to address the dynamics of opinion formation processes in social networks [29, 30]. The use of agent-based modeling offers the advantage that models can be used to simulate "how individuals and the environmental variables influencing them vary over space, time or other dimensions" [31, p. 11]. Thus, by generating models that represent agents and their interactions with a certain phenomenon under realistic conditions, it is possible to test even those theories that could otherwise only be addressed with very extensive empirical studies in which social interactions are observed in the long run [32]–[34]. It has been noted that there is a gap between the micro- (i.e., individual actions) and macro- (i.e., societal dynamics) level when it comes to investigating social phenomena [32, 35, 36].

Agent-based modeling studies also demonstrated the influence of OLs on the formation of opinions of individuals in their networks [25, 37, 38]. In a recent agent-based model, Borowski and colleagues [37] showed that the number of OLs and their ability to maximize information diffusion depends solely on the network structure. These results are in line with the study by van Eck and colleagues [25], who analyzed the influence of OLs and demonstrated that the velocity at which information is transmitted depends strongly on the network position of the OLs.

While these studies have corroborated the key role OLs play in social networks, they all modeled OLs as advocates for a certain stance who, in turn, were embedded in networks in which individuals were either supporting or opposing a political stance. However, social psychological research has consistently shown that holding an opinion on a political question can be more complex than just assuming a pro or contra stance [5]–[7].

Attitudinal ambivalence

Intuitively, most individuals would likely describe an opinion as something that can be roughly divided into the dichotomy of 'in favor' and 'against.' Similarly, research into political attitudes has, more or less implicitly, been conceiving of attitudes as one-dimensional constructs and measured them on scales that usually range between the two poles of 'completely against' and 'completely in favor.' One problem with this type of measurement is that responses often scatter closely around the scale's

midpoint, and researchers have been interpreting this as a 'neutral' attitude [3, 39]. However, this interpretation may not be valid, since attitudes are oftentimes more complex. The concept of 'attitudinal ambivalence' accounts for this complexity and describes attitudes as two-dimensional, i.e., it does not conceive of positive and negative evaluations of an attitudinal object as endpoints of one and the same continuum, but instead as two independent dimensions [3, 6, 40].

Accordingly, an ambivalent attitude simultaneously entails favorable and unfavorable evaluations toward an attitudinal object. As a consequence, a response on the scale midpoint may not only indicate neutrality, but instead be the result of an 'internal averaging' of the positive and negative evaluations [39]. Ambivalent attitudes are different from neutral attitudes or indifference (i.e., weak attitudes), since they entail equally strong evaluations of opposing poles [3]. In line with this, Thompson and colleagues [40] characterize ambivalent attitudes as being linked to equally strong positive and negative evaluations of at least moderate size. To determine the degree of ambivalence then, one simply needs to subtract the two attitude components from each other: the closer the resulting value is to 0, the higher the ambivalence toward the object (for a similar procedure, see [41]). Even though the inconsistency induced by ambivalence can be linked to aversive affective states in some cases [42], it is likely that ambivalent attitudes are, in general, very common: they have been found with regard to many different attitudinal objects, among them political issues (see [3, 43]). However, due to the widespread use of one-dimensional attitude conceptions in research on political opinions, they have likely been neglected in much of the extant work [39]. Ambivalent attitudes can promote conflicting intentions and thereby undermine the execution of behaviors linked to the attitudinal object (e.g., voicing an opinion in public; [44, 45]).

When it comes to discussions on social networking sites, those users who have an ambivalent attitude toward an issue may hence be less likely to express a clear stance toward that issue (and more likely to express balanced views). At the same time, ambivalence may result from exposure to political information in the first place [39, 45]. For instance, an individual might have a non-ambivalent attitude toward the COVID-19 policies but then get exposed to information supporting the contrary, thereby changing the overall evaluation of COVID-19 policies toward an ambivalent attitude. In short, attitudinal ambivalence can both shape the structure of an online discussion network and result from network effects. The present study takes both aspects into account and addresses the dynamics of mutual social influence in discussion networks that include the expression of ambivalent attitudes.

Psychological processes underlying attitudinal ambivalence

The observation that people's attitudes do not always fit into a unidimensional framework in the sense of being exclusively in favor or fully against something raises the question of how the prevalence of attitudinal ambivalence affects social influence dynamics in public opinion [4]. Being exposed to critical claims that oppose one's own is key to building mutual understanding and shared knowledge, as well as to fostering education for the effective functioning of democratic systems [46]. Following this logic, ambivalence may be desirable not only on a collective

(i.e., an ambivalent opinion climate among citizens) but also on an individual level (e.g., a person holding diametrical views on a certain political decision). Therefore, the present work focuses on the situation of being exposed to ambivalent rather than uniformly opposing views, the effects of which on political communication behavior is yet to be examined.

The missing link between ambivalence and political behavior prompts research to shed light on how attitudinal ambivalence manifests itself. Following the notion of the value pluralism model [47], scholars have proposed that attitudinal ambivalence should be associated with an "integrative complexity" or "balanced judgment," that is, individuals are capable of evaluating issues based on diverse and even contradictory information [48, 49]. While this state of ambivalence was found to evoke more thorough processing of newly incoming information [41, 50], ambivalent attitudes are more likely to increase uncertainty and induce more moderate attitudes [49]. Visser and Mirabile [51] demonstrated that people to whom individuals are directly connected, that is, in terms of their social networks, can be responsible for an individual's increasing attitudinal ambivalence. They argue that individuals compare their own attitudes with those of people around them and, in cases where they assess a conflict, they experience not only an intrapsychic tension due to the conflicting attitudes, but also "interpersonal conflictive tension" due to the connection to the person. Their findings show that attitudinally heterogeneous social networks (i.e., networks consisting of people with whom one agrees but also disagrees) foster individuals' attitudinal ambivalence and, therefore, decrease the strength of these attitudes. The power of the social environment and its influence on people's attitude, strength, and ambivalence has been emphasized and revealed by further studies [39, 52]. However, it remained unclear what role univalent and ambivalent key actors in networks—OLs—might play in the social influence process. So far, work on the effectiveness of influencers suggested that influential players in the network can have a significant impact on other users [53].

Method

Based on the outlined theoretical foundations and empirical evidence, we developed an agent-based model to examine our research questions (model, data, and results can be found in the repository of the Open Science Framework). For the implementation, we used NetLogo [54], which works in conjunction with the package RNetLogo [55].

Opinion domain

There are numerous models regarding the investigation of the evolution of opinions, and these models can be differentiated between discrete (Voter model: Clifford & Sudbury [56], Holley & Liggett [57], Sznajd model [58]) and continuous models (DeGroot [59], Deffuant-Weisbuch [60], Hegselmann-Krause [61]). While in the discrete models, the agents' value space is binary, in continuous models, it can be in

a continuous value interval. In continuous models, we can distinguish between those in which agents have their opinions influenced based on like-minded neighbors (bounded confidence) and those in which opinions are updated based on a weighted average of neighbor's opinions. In addition to these classical models, there are specific models that, although not developed primarily for opinion dynamics processes, are still suitable for this kind of dynamics process. The SIR (Susceptible, Infectious, or Recovered) model was first used to forecast the spread of diseases based on mathematical equations [62] and was also applied to opinion dynamics problems [63–65]. However, due to the different phases of this model, problems arise in the opinion consensus of the group, which is why further models for opinion dynamics processes were developed [66]. Furthermore, it was found that individual key nodes with a higher degree do not have a higher influence on the neighboring nodes [67]; these effects make the SIR model not further applicable for our consideration of OLs.

Our model, as specified below, is related to the DeGroot model, which follows the principle of social influence and assumes that the adjacent neighbors of an individual have an influence on opinion formation. In general, the DeGroot model implies that the agents strive for a common consensus, which is mainly achieved by the individual weighting of the agents, where this is constant and thus static over the entire course of the process. In the DeGroot model, the individual's updated opinion is simultaneously determined based on the confidence weights of the edges in the network and thus as a weighted average of their own current opinion and that of their neighbors [59]. One of the reasons we decided to focus on the DeGroot model is that it has been already used for numerous studies in the field of opinion dynamic due to its simple mechanism of updating opinions and its ease of extension, which allows researchers to customize the model according to their individual circumstances and specifically to their object of study. In our case, using the DeGroot model, we can directly represent the social influence of an agent in its network environment by measuring the perceived opinion climate of nodes based on their connected neighbors to update the opinions of the agent in a two-dimensional spectrum. Therefore, the use of the DeGroot model fits well to answer our research questions, as social influence and the related opinion dynamics can be considered in a social network. This is based on an iterative averaging model, in which agents' opinions are linked to neighboring nodes and thus considered to determine a kind of "opinion climate." This aspect of social influence is especially relevant for opinion leaders, as they operate in social network structures and can influence followers through their opinions. Since the DeGroot model is based on graph theory, the connections of nodes are enabled by means of edges and can be used for complex computations. Just as important as in the DeGroot model, the consideration of opinions from neighboring nodes has a major contribution in our implementation of the ABM to compute the perceived opinion climate and determine their update function. However, there are two major dissimilarities from the original DeGroot model that are manifested in our model. In our model, we have extended the DeGroot model to a two-dimensional opinion spectrum (an agent has a red opinion and a blue opinion), which are in competition with each other and allow agents to exhibit ambivalent behavior. By realizing the two-dimensional opinion observation, more complex

mechanisms can be captured. Previous studies have also shown that the results from a two-dimensional opinion range lead to the same results as in a one-dimensional model [68, 69] Therefore, we assume that our results of a two-dimensional vector compared to a one-dimensional vector might be similar for the first research question. Another difference to the original DeGroot model is the update function. Since this is static in the initial DeGroot model and does not change for the individual agents, we have introduced a dynamic gradation which depends on the individual opinion of the agent and the strength of the perceived opinion climate. This update function is applied to the two-dimensional opinion image of the individual agents and can change continuously as it progresses through the perceived opinion climate. To represent a detailed and transparent illustration of our model, we have depicted the sequential flow of our applied model in Fig. 1 below.

Figure 1 can be viewed in combination with the pseudocode from Table 1, which is responsible for the process of initializing the model, and Table 2, which contains the update function of the model. The graphical representation as well as

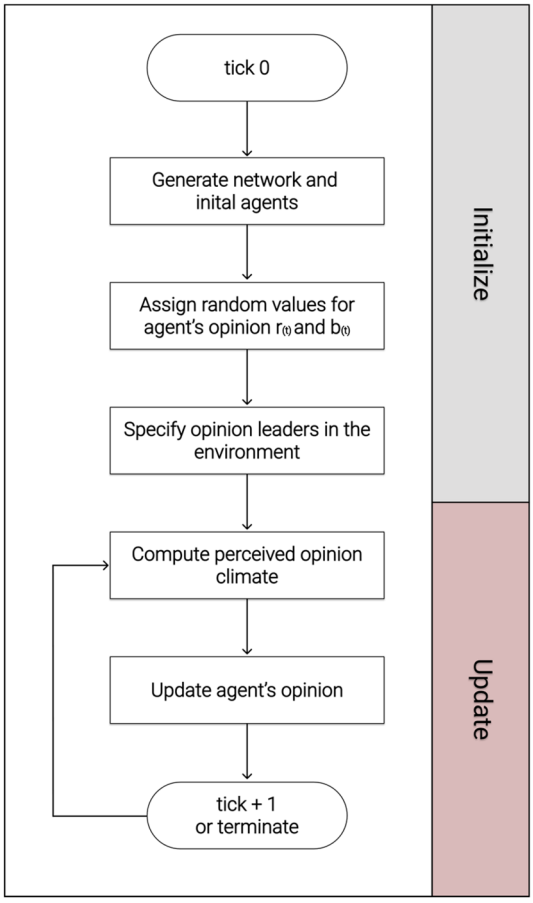

Fig. 1 Graphical representation of the model and its functionalities

Table 1 Mechanism and sequence of initialization of the model in pseudocode

AGENT-BASED MODEL: INITIALIZE			
1	**Generate network topologies and initial agents**		
1.1	*Set network type*		∇: *WS, PA*
1.2	*Set number of vertices (agents)*		V
2	**Assign random values for opinion red $r_{(i)}$ and opinion blue $b_{(i)}$**		

2.1	*For each agent* $\alpha \in V(G)$ {
	IF $r_{(i)} > b_{(i)}$
	THEN $\gamma_{color(i)} = red$
	IF $r_{(i)} < b_{(i)}$
	THEN $\gamma_{color(i)} = blue$
	IF $r_{(i)} = b_{(i)}$
	THEN $\gamma_{color(i)} = white$
	}
			R: [0, 1]

3	**Specify opinion leaders in the environment**		
3.1	*Set number of red, blue and ambivalent opinion leader*		σ_{red}, σ_{blue}, $\sigma_{ambivalent}$
3.2	*Assign opinion leadership to agents on degree centrality*		*Max. deg(*α*)*

3.2.1	*For each agent* $\alpha \in V(G)$ {
	Calculate degree of agents: deg($\boldsymbol{\alpha}$*)*
	
	According to pre-defined number of opinion leader, select agents with highest degree
	
	Update opinion values for opinion leaders:
	IF $\gamma_{color(i)} = red$ *AND* $\sigma_{red} > 0$
	THEN $r_{(i)} = 1$; $b_{(i)} = 0$
	IF $\gamma_{color(i)} = blue$ *AND* $\sigma_{blue} > 0$
	THEN $r_{(i)} = 0$; $b_{(i)} = 1$
	IF $\gamma_{color(i)} = white$ *AND* $\sigma_{ambivalent} > 0$
	THEN $r_{(i)} = [0, 0.5, 1]$; $b_{(i)} = [0, 0.5, 1]$
	
	Set isLeader = True
	}		

3.3	*Option of discrediting opinion leaders*		Ψ: True, False
3.3.1	*Set number of discrediting opinion leaders*		$-\sigma_{red}$, $-\sigma_{blue}$,
3.3.2	*Set negative value for discrediting the other opinion*		λ: [–0.1, –1]

3.3.3	*For the number of discredited opinion leader* $-\sigma_{red}$, σ_{blue} {
	Update opinion values
	IF σ_{red} = *isLeader AND* $r_{(i)} = 1$; $b_{(i)} = 0$
	THEN $b_{(i)} = \lambda_{red}$
	IF σ_{blue} = *isLeader AND* $r_{(i)} = 0$; $b_{(i)} = 1$
	THEN $r_{(i)} = \lambda_{blue}$
	}		

3.4	*Option of edge adjustment*		E: *True, False*
	Set total number of edges to rewire to red, blue and ambivalent opinion leaders *IF* E = *True AND* $\varepsilon_{red} > 0$ *THEN connect* σ_{red} *with random agent with low degree centrality* *IF* E = *True AND* $\varepsilon_{blue} > 0$ *THEN connect* σ_{blue} *with random agent with low degree centrality* *IF* E = *True AND* $\varepsilon_{ambivalent} > 0$ *THEN connect* $\sigma_{ambivalent}$ *with random agent with low degree centrality*		ε_{red}, ε_{blue}, $\varepsilon_{ambivalent}$

Table 2 Mechanism and sequence of the update function of the model in pseudocode

AGENT-BASED MODEL: UPDATE FUNCTION			
1	**Compute perceived opinion climate Ω_{red}, Ω_{blue}**		
	Get the degree of each agent in the neighborhood N(α)	*deg(α)*	
	Sum up red and blue opinions of each neighboring agent	$r\hat{}_{(t)}, b\hat{}_{(t)}$	
	Define perceived opinion climate by: $\Omega_{red} = \frac{\sum_{n=N(\alpha)} r\hat{}_{(t)}}{deg(\alpha)}$ $\Omega_{blue} = \frac{\sum_{n=N(\alpha)} b\hat{}_{(t)}}{deg(\alpha)}$		
2	**Update agents' opinion**		
2.1	*Set up ticks for runtime* *For i ← 0 to ticks*	*ticks*	
2.2	*Compute the difference between current opinion and the perceived opinion climate of the agent* $\delta_{red(t)} = r_{(t)} - \Omega_{red}$ $\delta_{blue(t)} = b_{(t)} - \Omega_{blue}$	$\delta_{red(t)}$, $\delta_{blue(t)}$	
3.3	*Adding an adjustment factor depending on* δ for red and blue opinion *A positive adjustment* $\Theta_{positive}$ *is given when:* *IF* $\delta < 0$ *AND* $\delta \geq -0.3$ *THEN 0.1* *IF* $\delta < -0.3$ *AND* $\delta \geq -0.7$ *THEN 0.2* *IF* $\delta < -0.7$ *THEN 0.3* *A negative adjustment* $\Theta_{negative}$ *is given when:* *IF* $\delta > 0$ *AND* $\delta \leq 0.3$ *THEN -0.1* *IF* $\delta > 0.3$ *AND* $\delta \leq 0.7$ *THEN -0.2* *IF* $\delta > 0.7$ *THEN -0.3*	$\Theta_{positive}$, $\Theta_{negative}$	
3.4	*Update the current opinion by recalculating the current opinion with the adjustment factor* $r_{(t)} \pm \Theta$ $b_{(t)} \pm \Theta$		

the pseudocode provide a deeper understanding to provide a more comprehensive understanding of the processes and their functionality. In the further course of the paper, the following sections implicitly refer to the description and explanation of the individual processes and how they are defined in detail.

Interaction direction and symmetry

We decided to represent the connection between individual agents and OLs in a network topology to guarantee the flow of opinions and their communication landscape. This means that the modeling is based on the principles of graph theory, where a graph G is defined as an ordered pair $G = (V, E)$, where V is a set of agents and E is a set of edges. We have adopted an undirected network, as this allows a bidirectional communication flow between agents, thereby ensuring that the agents not only express their own opinion but also come into direct contact with other opinions (thus resembling communication processes on social networking sites). Based on the theoretical background outlined earlier, we opted to apply two types of network topologies ∇ for our modeling: (a) the Barabási–Albert preferential attachment model [70] and (b) the Watts–Strogatz model [71]. Including both network topologies allowed us to test the robustness of the individual network structures. Since we build on a two-dimensional opinion model in our setup, we do not predict specific weights of the edges, but have used a predefined scheme to determine different thresholds that symmetrically update and adjust the two different opinions in the network.

Network models according to Barabási–Albert and Watts–Strogatz

Our investigated network structures have already been used in the literature based on opinion dynamics to explain specific use cases or social phenomena [34, 72]–[74]. In addition, the character properties emanating from the two network topologies are a crucial reason why we implemented them in our modeling in the context of opinion leadership. Here, the power law property is a key point, which is found not only in scale-free networks but has also been demonstrated in real social networks [75]. It was found that this distribution exists in both YouTube [76, 77] and Facebook [78] networks, where some users were also characterized with a very high degree. Based on our definition of OLs, we assume that social networking sites such as YouTube and Facebook provide a means for individuals to get in touch with other people and exchange opinions among themselves (for example, in the form of comments). In this process, OLs that have a significant impact on and influence the opinions of other users in the network can emerge. Furthermore, since this aspect in particular reflects the preferential attachment model, we decided to include the Watts–Strogatz model as a comparison in our analysis. Both network topologies are well-established models in different fields of science to explain complex structures and dynamic processes of networks in the real world and still have significant relevance to expand the understanding of network science today [79]. While the Watts–Strogatz model is based on a kind of friendship network, where friends are connected to other friends (clustered connectivity), the Preferential Attachment Model aims rather at the formation of individual hubs, which have more relationships to other nodes because they appear more attractive (heterogeneous connectivity). Furthermore, Hein et al. point out that network topologies may have an influence on the outcome of simulation studies, which makes it even more important to investigate different topologies [80]. Accordingly, examining these two network topologies in relation to opinion

leaders might reveal a way to infer statements about opinion dynamics and their influence on neighboring agents.

The Watts–Strogatz model is a randomized network belonging to the family of small-world networks that is more common in reality and is characterized by properties such as high clustering coefficients and short average path length [71]. Studies that have looked at information dissemination have also found that Watts–Strogatz networks perform similarly to scale-free networks [81, 82]. For these aforementioned reasons and our definition of OLs, we decided to use these two network topologies in the further course of our modeling. We describe the individual network models in more detail below:

The preferential attachment model by Barabási and Albert produces networks that are scale-free, i.e., have a power law degree distribution [70]. The actual functionality of network generation is that new nodes are preferentially connected to nodes with a high degree of connectivity, which ensures that an older node has many connections. Scale-free networks are based on the principle of preferential attachment and thus automatically provide a dynamic network structure, i.e., the addition of new nodes to an already well-connected node [37]. The equation of the preferential mechanism is defined by

$$P(k_i) = \frac{k_i}{\sum_j k_j}$$

where P is the probability to link a newly connected node to node i, which is dependent on the degree ki of node i. This mechanism results in a power law distribution and thus the principle "the rich get richer," where nodes with a high degree are preferred. Due to its natural ability to generate networks with power-level degree distributions, the Barabási–Albert model is generally deemed a good choice for modeling social networks.

Rewiring edges

To ensure that we fitted the definitions of OLs with numerous connections and greater influence on the opinion climate in the network, we decided to consider and apply the basic idea of randomly adding further edges to random OLs after the network had been created, whereby a low degree value was observed. The parameter (ϵ) can be enabled or disabled and assigns the number of edges that randomly connect from an agent to a random OL (ϵ_{blue},ϵ_{red},$\epsilon_{ambivalent}$). This rewiring makes it possible to create different distributions of OLs, where, for example, one faction has fewer OLs, but is very strongly connected, while the other faction is less strongly connected but has more OLs. Figure 2 shows the generated network topologies with connected agents and OLs.

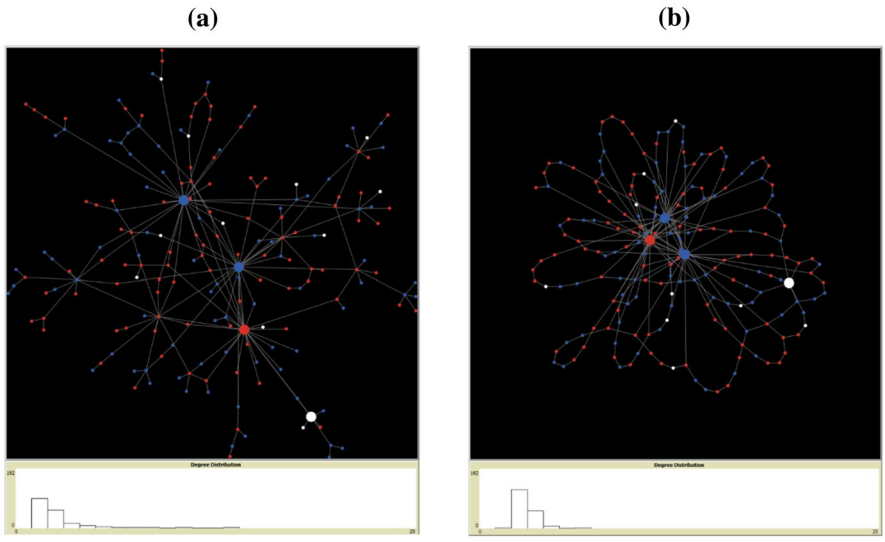

Fig. 2 Network topologies: (**a**) Barabási–Albert, (**b**) Watts–Strogatz

Interacting agents

In our model, each agent α holds a two-dimensional opinion, i.e., a "red" opinion $r_{(t)}$, and a "blue" opinion $b_{(t)}$, which represents a stance contrary to $r_{(t)}$ on the same issue. Opinions are represented by real values in the [0, 1] interval and change in several time intervals t. Values represent the extent to which a person is for (r) or against (b) a certain issue. It is important that this value is randomly assigned between 0 and 1, so that no bias occurs. To avoid very long decimal numbers, we round the value to the first decimal place. The value between 0 and 1 can be interpreted to some extent as possible arguments for the respective opinion. More precisely, if α has the initialization value of $r_{(0)} = 0.4$ and $b_{(0)} = 0.2$, it could be interpreted that it has four arguments in favor of the red stance, while it has only two arguments favoring the blue stance. These two values determine the attitude of each agent for the classification in a certain opinion camp as well as for the presentation in the network.

Thus, if the value of $r_{(t)}$ is higher than the value of $b_{(t)}$, this means that there are more arguments from the red opinion than from the blue opinion, and thus, the agent is inclined to favor the red opinion. If both opinions hold the same value, the agent holds an ambivalent opinion and is marked as a white node in the network. The color $\gamma_{color_{(t)}}$ represents the agent's opinions over time and is controlled in the model with the following rule:

$$\gamma_{color_{(t)}} = \begin{cases} red, & if\ r_{(t)} > b_{(t)} \\ blue, & if\ r_{(t)} < b_{(t)} \\ white, & if\ r_{(t)} = b_{(t)} \end{cases}$$

These values constantly change in the course of the simulation and, therefore, the properties of the agent in the network also change.

To determine the perceived neighboring opinion climate Ω of connected nodes from α, it is essential to identify $\widehat{r_{(t)}}$ and $\widehat{b_{(t)}}$ of the neighborhood $N()$. The neighborhood of an agent $\alpha \in V(G)$ is the set of all nodes that are adjacent to α and defined as $N(\alpha) = \{y \in V(G) : \{\alpha, y\} \in E(G)\}$. Starting with one agent, we calculate the neighborhood and determine the sum of $\widehat{r_{(t)}}$ and $\widehat{b_{(t)}}$ of all connected agents in the network. The sum of this is the perceived neighboring opinion climate of a single agentα. To include the number of connected nodes in the calculation of the perceived neighboring opinion, this is divided by the degree of alpha $deg(\alpha)$. The degree is defined as the number of nodes adjacent to α and therefore the size of the neighborhood of α, that is, $deg(\alpha) = |N(\alpha)|$. The following equations show the perceived neighboring opinion climate of a specific agent and their neighboring opinions $\widehat{r_{(t)}}$ and $\widehat{b_{(t)}}$:

$$\Omega_{red} = \frac{\sum\limits_{n=N(\alpha)} \widehat{r_{(t)}}}{deg(\alpha)}$$

$$\Omega_{blue} = \frac{\sum\limits_{n=N(\alpha)} \widehat{b_{(t)}}}{deg(\alpha)}$$

With the newly computed factor of the perceived neighboring opinion climate Ω_{red} and Ω_{blue}, the difference $\delta_{red(t)}$ and $\delta_{blue(t)}$ to the outgoing single agent opinions $r_{(t)}$ and $b_{(t)}$ is then determined. This value is rounded to the third decimal place and allows for a more fine-grained view of opinions and the further course. This computation is necessary, since it ensures that the existing opinion climate of an agent is directly referenced with their own opinion, thus taking into account the effect of the network and its nodes.

$$\delta_{red(t)} = r_{(t)} - \Omega_{red}$$

$$\delta_{blue(t)} = b_{(t)} - \Omega_{blue}$$

Updating function

If the value of the differential is positive, a negative adjustment factor $\Theta_{negative}$ is added to the current opinion, while for a negative value of the differential, a positive adjustment factor $\Theta_{positive}$ is updated to the current opinion. Thus, we ensure with the following formulas, $r_{(t)} \pm \Theta$ and $b_{(t)} \pm \Theta$, that the two opinions are recalculated per tick to guarantee a constant adaptation of the model. Finally, we also round the value of the newly computed opinion to the second decimal place, since the implementation of an ambivalent opinion space ($r_{(t)} == b_{(t)}$) is otherwise not feasible due to an extremely small probability and doing so also improves the performance within our model for the further course. We have chosen the following adjustment values Θ for specific intervals, so that even in a more strongly represented opinion climate, the adaptation of the

individual agent manifests itself in a stronger form. To be more precise, if δ is greater than the agent's opinion, this means that the agent adapts their opinion to the climate of opinion. The stronger δ, and thus the perceived climate of opinion, the stronger the agent adapts to this opinion—and likewise, even if the perceived climate of opinion is represented as weaker than the current opinion. Here, the value of the opinions is then corrected downward, since the agent feels no influence of their environment.

$$\Theta_{positive} = \begin{cases} 0.1, \ if \ \ \delta < 0 \ and \ \delta \geq -0.3 \\ 0.2, \ if \ \ \delta < -0.3 \ and \ \delta \geq -0.7 \\ 0.3, \ if \ \ \delta < -0.7 \end{cases}$$

$$\Theta_{negative} = \begin{cases} -0.1, \ if \ \ \delta > 0 \ and \ \delta \leq 0.3 \\ -0.2, \ if \ \ \delta > 0.3 \ and \ \delta \leq 0.7 \\ -0.3, \ if \ \ \delta > 0.7 \end{cases}$$

This approach to modeling conceives of influence as a linear function of opinions within the social environment, as found in various accounts of opinion formation and social influence [27, 28, 83] and as previously found in large-scale social network data [17]. While work on psychological reactance and so-called "backfire effects" [84, 85] would suggest a more complex process that includes the possibility of non-linear adaptions (e.g., opinion change in the opposite direction of opinions represented within the network environment), the present work models opinion dynamics based on accounts that are more parsimonious (yet empirically well-founded) and which allow our model to remain more simple.

Interacting opinion leaders

In our model, OLs σ have the same characteristics as other agents (i.e., users), but they differ in their position in the network and in the constant opinion values with which they influence the opinions of connected agents. Since the position of OLs is a key factor for the diffusion of information in the network [25], we adhered to the results of previous research in our modeling. There are different approaches in terms of centrality measures (degree, betweenness, and closeness) to determine these nodes of OLs in networks. Xiao and colleagues [15] also used an agent-based model to investigate the dynamic processes of OLs in networks and found that the detection of the three types of centralities (degree centrality, betweenness centrality, and closeness centrality) have a similar influence on opinion formation and thus differ only marginally. As previous research has demonstrated, OLs in networks are characterized by the fact that they hold a higher in-degree centrality in the network and therefore have more influence on individuals [15]. Other studies have also taken the measurement of degree centrality as a criterion for detecting OLs in real-world social networks ([86, 87]) and defined this as an indicator of local OLs [88, 89]. Once the network had been generated and each agent had been initialized, we characterized OLs on the basis of degree centrality. Degree centrality measures

the number of connections of nodes connected to a particular node; the higher the degree centrality of a node, the more influence it has in the network. The calculation of degree centrality is defined as follows:

$$C_{Di} = \sum_{j} a_{ij}$$

At this point, we only know which nodes have greater influence in the network; however, these OLs are not yet assigned to an opinion camp. The next step is to randomly assign the OLs; to ensure this, we compute the sum of the requested OLs across all opinion camps to filter and select only nodes characterized with the Nth highest degree centrality. Once we had identified the nodes with the highest degree centralities, the OLs were randomly assigned to the red and blue opinion according to the highest degree centrality of each node. The randomization process ensures that the distribution of the opinion camp is equitably distributed and that we do not create bias in the positioning of OLs in the network. This procedure also applies to the ambivalent OLs.

Univalent and ambivalent OLs

We distinguish between univalent and ambivalent OLs that are able to influence the opinions of the agents in the network. For the univalent OLs the values are different (i.e., $\sigma_{red} : r=1$ and $b=0$, or, $\sigma_{blue} : r=0$ and $b=1$) to model a strong univalent influence on the opinions. For ambivalent OLs, the values are identical and moderate (i.e., $\sigma_{ambivalent} : r=0.5$ and $b=0.5$) or high (i.e., $r=1$ and $b=1$). Consequently, OLs can equally influence both opinions of the agents they are connected with. Additionally, our modeling allows us to vary the total number of OLs in the network and thus to split the distribution specifically. Different distributions of OLs can be simulated in the model with regard to the number of ambivalent, red, or blue OLs present in the network.

Discrediting OLs

In addition to the functionality of the OLs, we have also considered the case of OLs that discredit their opponent's position. The functionality of discrediting Ψ can be switched on and off depending on the configuration of the model, where on the one hand the quantity of discrediting OLs $-\sigma$ and on the other hand the intensity of the discrediting λ toward the opposing opinion camp can be determined. The intensity of the discrediting λ has been implemented in our model by a negative value of [–0.1 to –1] for the opposite opinion, so that, e.g., an OL in the red opinion camp has the values $r=1$, $b=\lambda_{blue}$, $\lambda_{blue}=-0.5$. The negative values have a direct influence on the temporal course of the modeling, since OLs in this case not only increase neighboring agents' values with regard to one camp (e.g., $r_{(t)}$), but also decrease their values regarding the stance toward the opposing camp (e.g., $b_{(t)}$).

Figure 3 shows the mechanism of the influence of an OL and how the opinion climate changes. It can be seen that the blue OL has a strong influence on the ambivalent node as well as on the directly connected red node. Consequently, the OL has succeeded in influencing agents in their direct environment toward a majority of blue nodes.

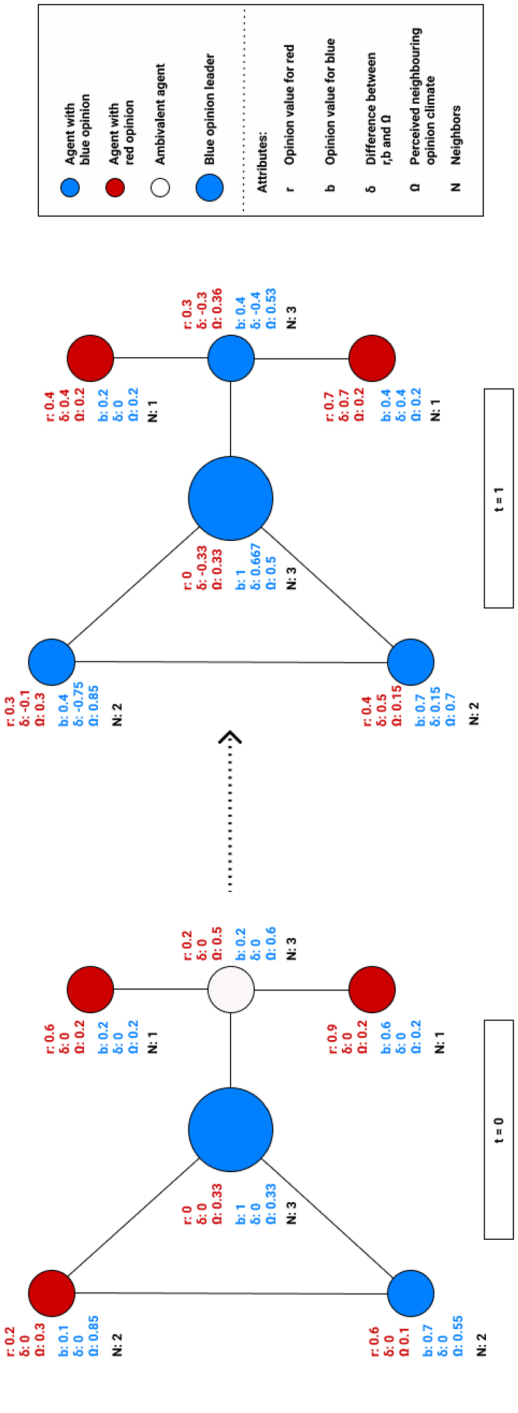

Fig. 3 Example of an opinion update

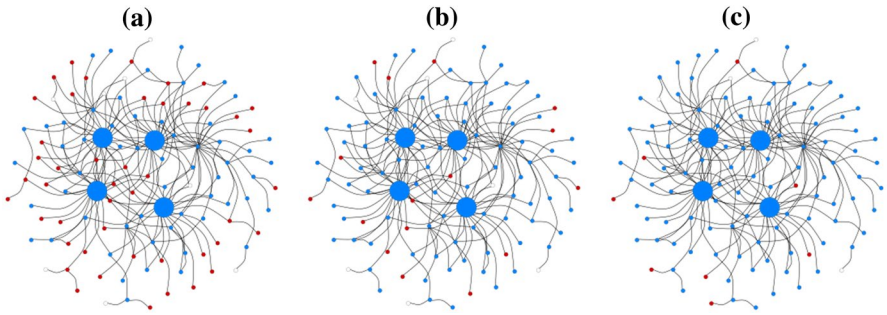

Fig. 4 Example of opinion formation by four opinion leaders: (**a**) at the beginning with ($\gamma_{blue} = 61$, $\gamma_{red} = 46$, $\gamma_{white} = 8$), (**b**) after eight ticks ($\gamma_{blue} = 91$, $\gamma_{red} = 16$, $\gamma_{white} = 8$), and (**c**) after 29 ticks ($\gamma_{blue} = 102$, $\gamma_{red} = 8$, $\gamma_{white} = 5$)

Figure 4 shows three examples of our generated networks of how the opinion climate regarding four OLs changes over several ticks.

Validation of the agent-based model

The validation of our agent-based model was carried out with a sensitivity analysis, i.e., a variation of different dimensions of input/output parameters, to determine how the parameter settings affect the behavior of the model. With the help of a sensitivity analysis, the different parameters of a model can be examined to increase the accuracy of the model, reduce the output variance, and simplify the model [90]. The various parameter spaces for our three scenarios to answer the three research questions are shown in Table 3. Here, the distribution of opinion leaders for the different opinion camps is particularly relevant. In combination with the total number of nodes including the number of opinion leaders in the network, we have adopted the divergent outcomes from the previous studies [2, 24] in our modeling and therefore implemented varying distributions of OLs in the network. We decided to consider 21 different distributions of opinion leaders, where in some distributions, a maximum of 20% of opinion leaders at 500 nodes and 10% of opinion leaders at 1000 nodes are investigated. Due to the variation in the distribution of OLs, the direct comparison between the ratio of opinion leaders from the different camps allows us to draw conclusions about the extent of influence on opinion formation in the network. Here, the two networks used can also ensure further conclusions about the topology in the network. Using the additional option of edge matching by means of a fixed number of randomized edges to the opinion leaders, we can clearly match the general definition of OLs, since they interact as a hub and are connected to many other agents in the network. The stepwise values for opinion leaders who take a discrediting opinion toward the other camp were chosen for the reason to be able to identify and compare possible tendencies of the opinion distribution more easily.

Furthermore, we applied the one-parameter-at-a-time (OAT) method to isolate each of the individual parameters and model it with different variations of other parameters. According to Lee et al. [91], this method increases the robustness of

Table 3 Summary of agent-based modeling parameters

Parameter	Explanation	Description	Parameter space
V	Nodes	Tells the total number of nodes in the network (OLs and normal agents)	500, 1000
E	Edge adjustment	Provides an edge control functionality to determine whether additional edges of random nodes are connected to the OLs, thus strengthening the effect of the OLs in the network	True
Ψ	Discrediting	Provides a discrediting functionality for OLs	True, false
∇	Network topology	The modeling can be executed considering the two network structures of preferential attachment or Watts–Strogatz	Preferential attachment, Watts–Strogatz
σ_{blue}, σ_{red}, $\sigma_{ambivalent}$	Number of blue, red, and ambivalent OLs	Represents the number of blue, red, and ambivalent OLs used in our modeling	[0, 1, 5, 12, 25, 50] [0, 1, 5, 12, 25, 50] [0, 1, 12, 20, 25, 50]
ϵ_{blue}, ϵ_{red}, $\epsilon_{ambivalent}$	Number of random connected edges to OLs	If E equals true, then the number of edges from random agents is randomly connected to OLs. Allows the modification of network structures and settings	100, 100, 100
$-\sigma_{blue}$ $-\sigma_{red}$	Number of discrediting OLs	Indicates the number of discrediting OLs (parameter can be greater than the actual number of OLs σ for the particular opinion camp)	[0] [0, 1, 5, 12, 25, 50]
λ_{blue} λ_{red}	Negative value for discrediting the other opinion	Represents a negative opinion value, which is in a range of –0.1 to –1. The higher the value, the higher is the negative influence in the network	0 [0, –0.2, –0.4, –0.6, –0.8, –1]

OLs opinion leaders

For each research question, we specified a different parameter space, so that different aspects such as baseline, ambivalent OLs, and discrediting OLs could be examined. In Appendix A (supplementary material), a more detailed subdivision of the parameter spaces has been included, which addresses the individual research questions

the model with repeated iterations. For our modeling, we set a value of 1000 for the iteration of each parameter space. As Waldherr & Wettstein [32] suggest, we determined an average result based on multiple iterations of the same parameter space. The variously applied parameter settings are not only intended to make the model realistic and to test it in this respect, but also to identify how the model behaves, for example at extreme values. We decided to set the number of agents in both networks based on previous studies that used agent-based modeling to investigate different social science theories such as the spiral of silence theory [10, 73], opinion leaders [15, 92], or general opinion dynamics [93, 94].

Results and simulation experiments

To address our research questions, we generated an agent-based approach with 21 different distributions of OLs (see Table 3), varying with regard to the representation of the two opinion groups. Based on the different distributions, we can determine the ways in which the overall opinion climate is affected (e.g., the circumstances under which one OL gains a majority).

RQ1 asked how the opinion climate responds to a varying ratio of OLs who promote opposing stances on the same issue. To address RQ1, we first tested the scenario in which no OLs are present. Due to the fact that the modeling for the various parameter settings was performed on 1000 iterations, the results are considered on the averaged final state shown in Table 4. Here, on average, based on both network topologies, 5.48% of the nodes are ambivalent, 23.07% belong to the minority group, and slightly more than 71.44% are in the majority group. However, the subdivision of the two network topologies reveals differences in the distribution of the opinion climate. The most pronounced differences with regard to the applied network topology appear in the distribution of the ambivalent nodes when zero OLs are present in the network: while in the preferential attachment topology, 17.48% of nodes are ambivalent, this is the case for only 3.2% in the Watts–Strogatz network. Similarly, differences can be found in the minority group when there are no OLs: while this proportion is 31.53% in the preferential attachment network, it is 42.66% in the Watts–Strogatz network. The results for the majority group differ only slightly (50.99% preferential attachment, 54.14% Watts–Strogatz). However, the results for a strongly unbalanced distribution of OLs (i.e., 0 vs. 12 and 0 vs. 25, 0 vs. 50) demonstrate that in both network topologies, the majority group has a 93% win rate. Furthermore, when a critical mass of OLs is reached (at least 12 in each group), the fact that one opinion has twice as many OLs leads to further changes in the distribution of opinion climate.

To make further statements about RQ1 with respect to the ambivalent nodes and changes in the opinion climate, we compared the initial and final shares of ambivalent nodes present in the network. This allows us to make a statement about the number of ambivalent nodes at the initialization of the network and at the end of the modeling. Interestingly, there were also differences in the ratios of the preferential attachment and Watts–Strogatz models when no OLs are present (see Fig. 5). In a state without OLs, the findings show that in a preferential attachment network,

Table 4 Results of the modeling with different opinion leader (OL) distribution and network topologies to evaluate the opinion distribution

OL distribution	Network topology	% Majority	% Minority	% Ambivalent
0	PA	50.99 CI [510.57, 51.4]	31.53 CI [31.21, 31.85]	17.48 CI [17.12, 17.85]
	WS	54.14 CI [54.34, 54.94]	42.66 CI [42.46, 42.86]	3.2 CI [3.13, 3.27]
0 vs. 1	PA	74.37 CI [74.08, 74.66]	18.01 CI [17.81, 18.22]	7.62 CI [7.48, 7.75]
	WS	75.64 CI [75.45, 75.83]	21.55 CI [21.37, 21.73]	2.81 CI [2.79, 2.84]
0 vs. 5	PA	90.12 CI [90.02, 90.21]	8.24 CI [8.17, 8.32]	1.64 CI [1.61, 1.67]
	WS	89.97 CI [89.88, 90.05]	8.39 CI [8.32, 8.46]	1.64 CI [1.62, 1.66]
0 vs. 12	PA	93.07 CI [93.02, 93.13]	6.14 CI [6.09, 6.18]	0.79 CI [0.78, 0.81]
	WS	92.8 CI [92.75, 92.86]	6.18 CI [6.14, 6.23]	1.01 CI [1, 1.03]
0 vs. 25	PA	93.24 CI [93.19, 93.29]	6.03 CI [5.98, 6.07]	0.74 CI [0.72, 0.75]
	WS	92.89 CI [92.84, 92.94]	6.13 CI [6.09, 6.18]	0.98 CI [0.97, 1]
0 vs. 50	PA	93.36 CI [93.31, 93.41]	5.97 CI [5.93, 6.01]	0.67 CI [0.66, 0.68]
	WS	92.99 CI [92.94, 93.04]	6.07 CI [6.03, 6.11]	0.94 CI [0.93, 0.96]
1 vs. 1	PA	48.5 CI [48.35, 48.66]	40.47 CI [40.33, 40.62]	11.02 CI [10.89, 11.16]
	WS	49.27 CI [49.18, 49.36]	44.07 CI [43.98, 44.16]	6.66 CI [6.59, 6.73]
1 vs. 5	PA	78.41 CI [78.32, 78.5]	16.78 CI [16.7, 16.86]	4.81 CI [4.77, 4.85]
	WS	78.63 CI [78.53, 78.73]	16.51 CI [16.42, 16.59]	4.86 CI [4.84, 4.89]
1 vs. 12	PA	88.09 CI [88.03, 88.16]	9.54 CI [9.48, 9.6]	2.37 CI [2.34, 2.39]
	WS	87.8 CI [87.72, 87.87]	9.42 CI [9.36, 9.48]	2.78 CI [2.76, 2.81]
1 vs. 25	PA	88.64 CI [88.58, 88.71]	9.18 CI [9.12, 9.23]	2.18 CI [2.16, 2.2]
	WS	88.16 CI [88.08, 88.23]	9.17 CI [9.11, 9.23]	2.68 CI [2.66, 2.7]
1 vs. 50	PA	89.04 CI [88.98, 89.1]	8.91 CI [8.86, 8.96]	2.05 CI [2.03, 2.07]
	WS	88.63 CI [88.56, 88.7]	8.88 CI [8.82, 8.94]	2.49 CI [2.52, 2.47]

Table 4 (continued)

OL distribution	Network topology	% Majority	% Minority	% Ambivalent
5 vs. 5	PA	46.97 CI [46.9, 47.03]	43.75 CI [43.69, 43.81]	9.29 CI [9.22, 9.35]
	WS	46.78 CI [46.72, 46.84]	43.71 CI [43.65, 43.77]	9.51 CI [9.44, 9.57]
5 vs. 12	PA	67.47 CI [67.4, 67.54]	25.14 CI [25.07, 25.2]	7.39 CI [7.36, 7.43]
	WS	65.55 CI [65.48, 65.61]	26.05 CI [25.98, 26.13]	8.4 CI [8.36, 8.44]
5 vs. 25	PA	69.88 CI [69.81, 69.95]	23.33 CI [23.27, 23.4]	6.79 CI [6.76, 6.82]
	WS	67.02 CI [66.95, 67.1]	24.97 CI [24.89, 25.04]	8.01 CI [7.98, 8.04]
5 vs. 50	PA	71.65 CI [71.58, 71.73]	21.98 CI [21.92, 22.05]	6.36 CI [6.33, 6.4]
	WS	69 CI [68.91, 69.08]	23.52 CI [23.44, 23.6]	7.48 CI [7.45, 7.51]
12 vs. 12	PA	46.86 CI [46.8, 46.92]	43.7 CI [43.64, 43.76]	9.44 CI [9.38, 9.5]
	WS	46.17 CI [46.11, 46.22]	43.32 CI [43.26, 43.38]	10.51 CI [10.45, 10.57]
12 vs. 25	PA	49.05 CI [48.98, 49.11]	41.97 CI [41.9, 42.03]	8.99 CI [8.95, 9.03]
	WS	47.31 CI [47.26, 47.36]	42.65 CI [42.59, 42.7]	10.04 CI [10, 10.08]
12 vs. 50	PA	52.2 CI [52.13, 52.27]	39.37 CI [39.3, 39.43]	8.44 CI [8.4, 8.47]
	WS	50.11 CI [50.05, 50.18]	40.39 CI [40.32, 40.46]	9.49 CI [9.46, 9.53]
25 vs. 25	PA	47.68 CI [47.61, 47.75]	43.82 CI [43.75, 43.89]	8.5 CI [8.45, 8.55]
	WS	46.52 CI [46.46, 46.57]	43.82 CI [43.76, 43.87]	9.67 CI [9,61, 9.72]
25 vs. 50	PA	49.22 CI [49.16, 49.29]	42.73 CI [42.66, 42.79]	8.05 CI [8.02, 8.09]
	WS	48.23 CI [48.17, 48.28]	42.64 CI [42.58, 42.69]	9.14 CI [9.1, 9.17]
50 vs. 50	PA	48.06 CI [48, 48.13]	44.32 CI [44.26, 44.39]	7.61 CI [7.57, 7.66]
	WS	47.02 CI [46.97, 47.07]	44.37 CI [44.32, 44.42]	8.61 CI [8.56, 8.66]
Average	Both	71.44 CI [71.35, 71.54]	23.07 CI [22.99, 23.15]	5.48 CI [5.46, 5.50]

the number of ambivalent nodes in the network increases [from 9.49% CI (9.44%, 9.54%) to 17.48% CI (17.12%, 17.85%)], while in the Watts–Strogatz architecture, the number of ambivalent nodes decreases [from 9.46% CI (9.41%, 9.51%) to 3.20% CI (3.13%, 3.27%)]. However, this scenario without any OLs is relatively unlikely. Furthermore, it again appears that with a critical mass of OLs present in the network, the share of ambivalent opinions differs only by a minimal percentage. Our findings show that the number of ambivalent nodes remains comparatively constant over time, unless there is a strong imbalance in OLs (0:5, 0:12, 0:25 or 0:50).

As soon as an imbalance arises (0:12 or 0:25), the opinion camp with the higher number of OLs always wins. As soon as there are at least 12 OLs in both camps, the number of OLs does not matter, and the opinion climate becomes more balanced again. Figure 6 demonstrates the normalized opinion distribution of ambivalent agents at the initial state of the model and after the termination of the model and is clustered based on the two network topologies. The percentage distribution of ambivalent opinions can be seen on the y-axis, while the x-axis considers different distributions of OLs in the network.

RQ2 asked how the distribution of opinion climate responds to OLs who advocate in favor of one stance (univalent OLs) compared to ambivalent OLs. To this end, we ran a further model, this time including a varying number of ambivalent OLs (0, 1, 12, 20, 25, 50). This scenario thus includes OLs who advocate for the two opposing stances r and b with equal strength (i.e., $r=0.5$ and $b=0.5$ *or* $r=1$ and $b=1$). Furthermore, to increase the realism of the model, we decided to additionally include univalent OLs who represent the red or blue opinion camp. For this reason, we have included different numbers of OLs in the model over several iterations. Figure 7 shows a normalized opinion distribution based on the grouping of the ambivalent opinion values (0.5 vs. 0.5; 1 vs. 1), which include OLs and the resulting percentage distribution of the opinion climate (y-axis), separately for the two network topologies. The x-axis shows the number of OLs in the network (the results are averaged over multiple iterations).

When OLs are ambivalent, the results show that the number of agents with ambivalent opinions increases in the network. The difference between the two network topologies is very small and differs only marginally. The greatest effect is revealed when ambivalent moderate opinion strength with equally strong opinions of red and blue opinions (0.5 vs 0.5) is expressed in the network; an average percentage value of ambivalent agents of 11.40% CI [11.38%, 11.42%] after the end of the modeling results. Furthermore, in this context, it is noticeable that 1 OL and 12 OLs in the preferential attachment network have a percentage of 7.03% CI [6.98%, 7.08%] and 12.59% CI [12.52%, 12.65%] ambivalent agents, while 20, 25, and 50 OLs have a percentage of 13.32% CI [13.25%, 13.38%], 13.72% CI [13.66%, 13.79%], and 15.68% CI [15.62%, 15.74%] ambivalent agents. However, with a distribution of 1 versus 1, the average value of the climate of opinion in relation to the ambivalent nodes is lower at a value of 8.79% CI [8.78%, 8.80%]. Having said that, it is also shown that with an increase in OLs, the climate of opinion becomes more ambivalent. For example, the value is 10.87% CI [10.84%, 10.90%] for 12 OLs and rises to 14.11% CI [14.09%, 14.14%] for 50 OLs.

In addition to the pure number of ambivalent OLs and their influence, the interaction between the different OLs is a key point to better understanding the dynamics of opinion formation. For this reason, we took a closer look at the climate of opinion on the different combinations of OLs in the network. More precisely, we examined the interaction and thus also the influence of the ambivalent OLs in relation to other OLs belonging to a certain opinion camp. In our model, as well as in the consideration of the results, we assume different distributions of the OLs (red OLs vs. blue OLs vs. ambivalent OLs). As an example, the following distribution would mean that there are no OLs in the red and blue opinion camps, but only 12 ambivalent OLs. Figure 8 shows this interaction between the different OLs of the opinion camps and the resulting distribution of opinion climate.

We exclusively focused on these ambivalent opinion values in the analysis, since, according to our definition, only the ambivalent OLs are considered to obtain further information.

The results show that in a distribution with only ambivalent OLs and a value of $r=0.5$ and $b=0.5$ for the opinions, the climate of opinion becomes balanced (35%). In this case, the climate of opinion of red and blue agents in the network is identical. With stronger opinion values of $r=1$ and $b=1$, ambivalent opinions in the network tend to decrease, and instead, red and blue opinions are more present in the network (42%). The results also show that if there is an imbalance of OLs in one opinion camp (0 vs. 12 vs. 25), it is not possible for those who are ambivalent OLs to create a balanced opinion climate. In Table 5, only a small percentage of ambivalent opinions (5–10%) and minority opinions (red, 8–11%) is shown, while the blue camp with the 12 OLs reaches a share of 79–87% in the network. The results in Table 5 show that, compared to the distribution of 12 versus 12 versus 12, there is still a relatively similar number of ambivalent agents. In particular, the distribution of agents in the blue opinion camp is then 43% for the preferential attachment and for the Watts–Strogatz model.

With regard to RQ2, it can be summarized that ambivalent OLs can ensure that the opinion climate concerning red and blue opinion is relatively stabilized and thus ultimately contribute to the existence of more ambivalent agents. Ambivalent OLs have the greatest influence when there are no other OLs from specific opinion camps in the network.

RQ 3 asked how the distribution of opinion climate responds to univalent OLs who discredit their opponent's stance. To investigate the effect of discrediting OLs, we investigated five different distributions of OLs (1 vs. 1, 5 vs. 5, 12 vs.12, 25 vs. 25, and 50 vs. 50) in each camp r and b, since this scenario from the above findings leads to a stable opinion climate in which the two opinion camps are relatively equally distributed. Furthermore, we decided to increase the negative discredit value by –0.2 steps until the value of –1 was reached. In this scenario, we assumed that only OLs in the red camp may take on the role of discrediting the blue camp to have a clear comparative value. Regarding RQ3, the results in Fig. 9 show that the stronger the discrediting expression of an OL, the higher the probability that the opinion climate tips over in favor of the OL's camp, in contrast to which supporters of the other party shrink to a minority. These findings can be found in both networks, the preferential attachment and the Watts–Strogatz network architecture.

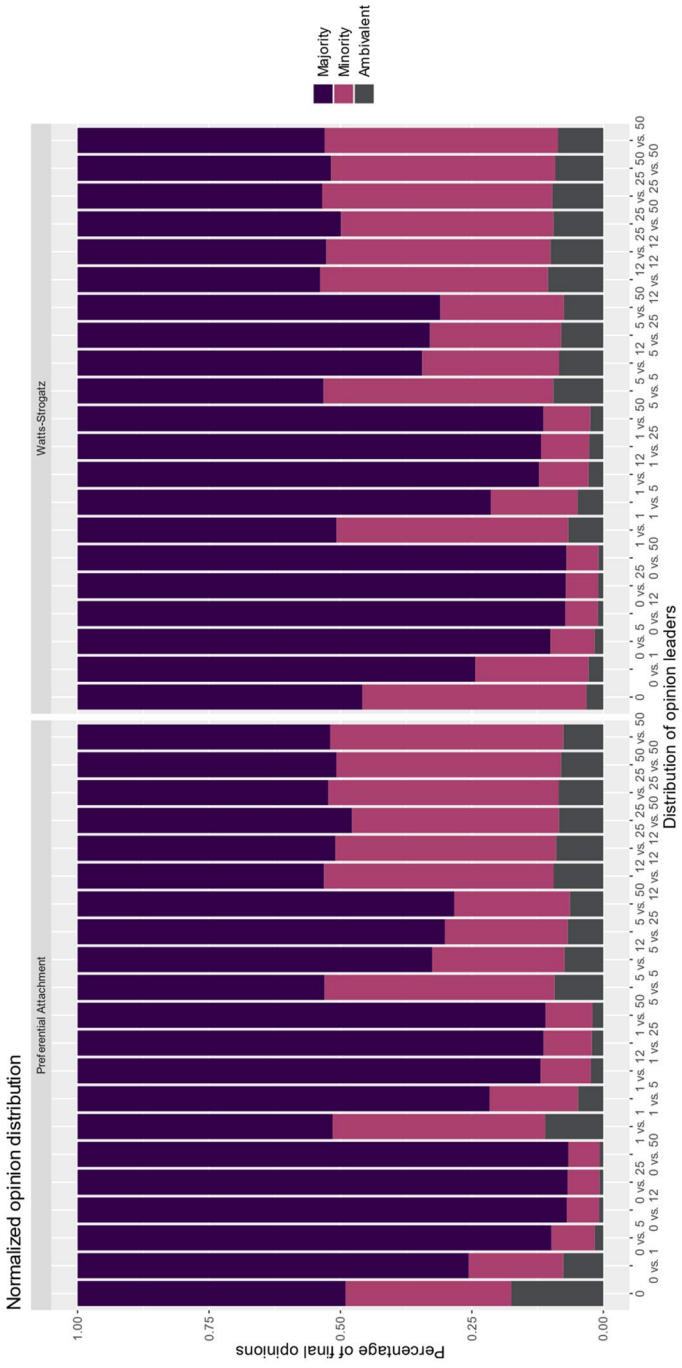

Fig. 5 Normalized opinion distribution of 'majority,' 'minority,' and 'ambivalent' in relation to the distribution of opinion leaders

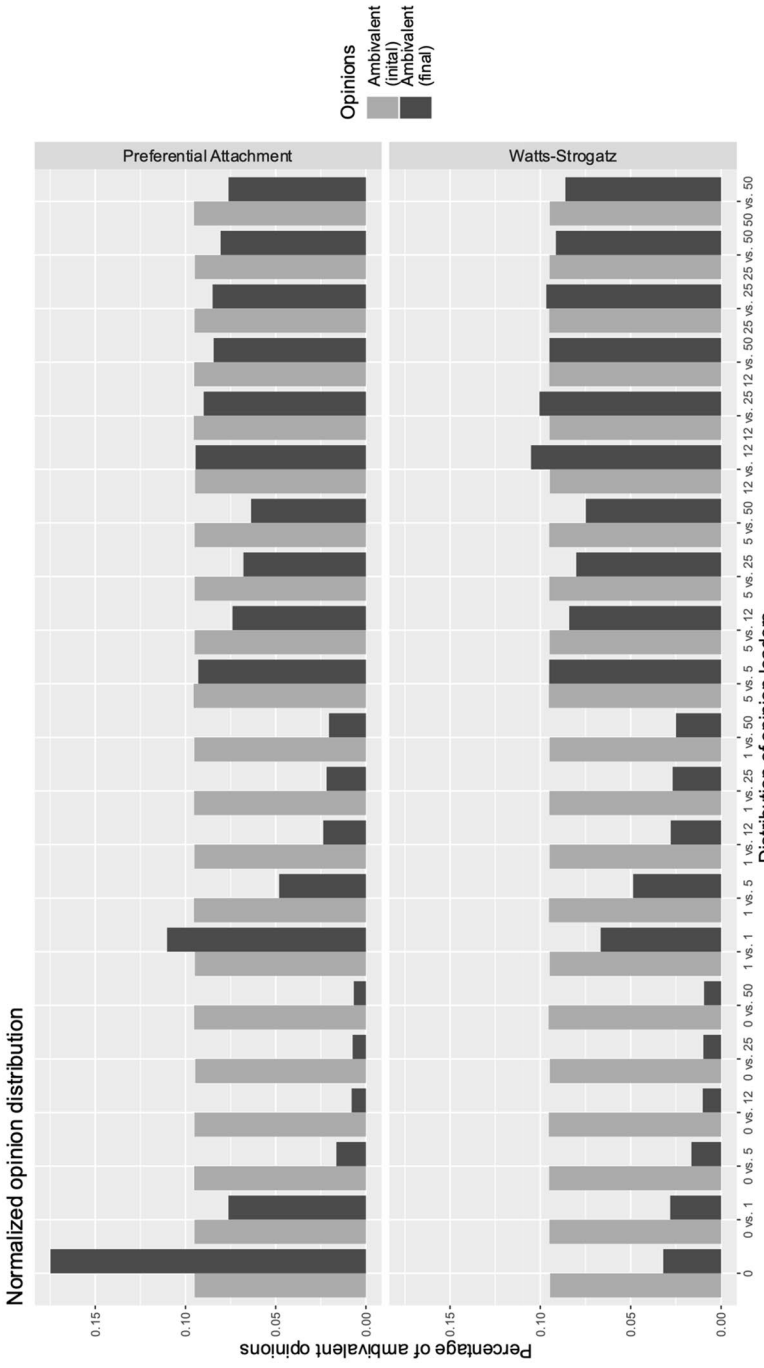

Fig. 6 Results of the initial and final values of the ambivalent nodes in relation to the distribution of opinion leaders

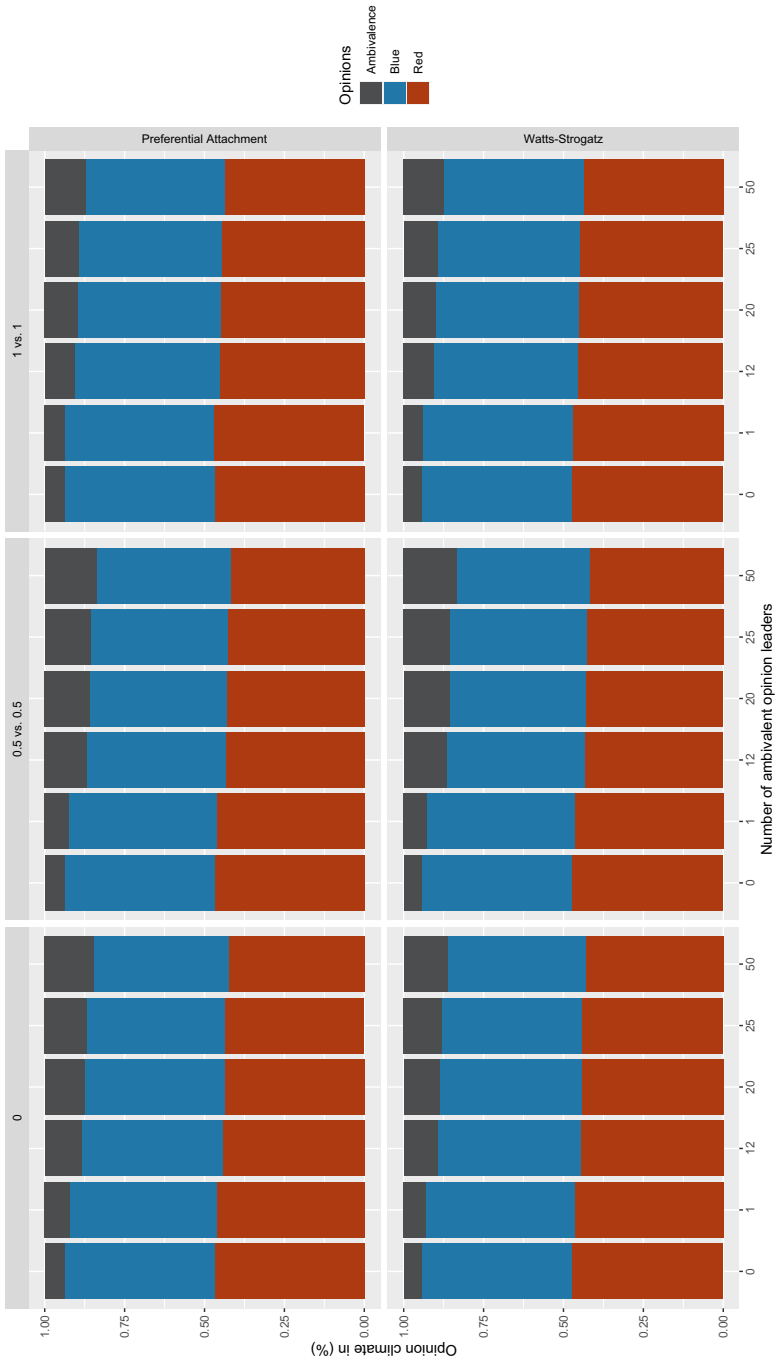

Fig. 7 Impact of ambivalent opinion leaders on the climate of opinion in the network

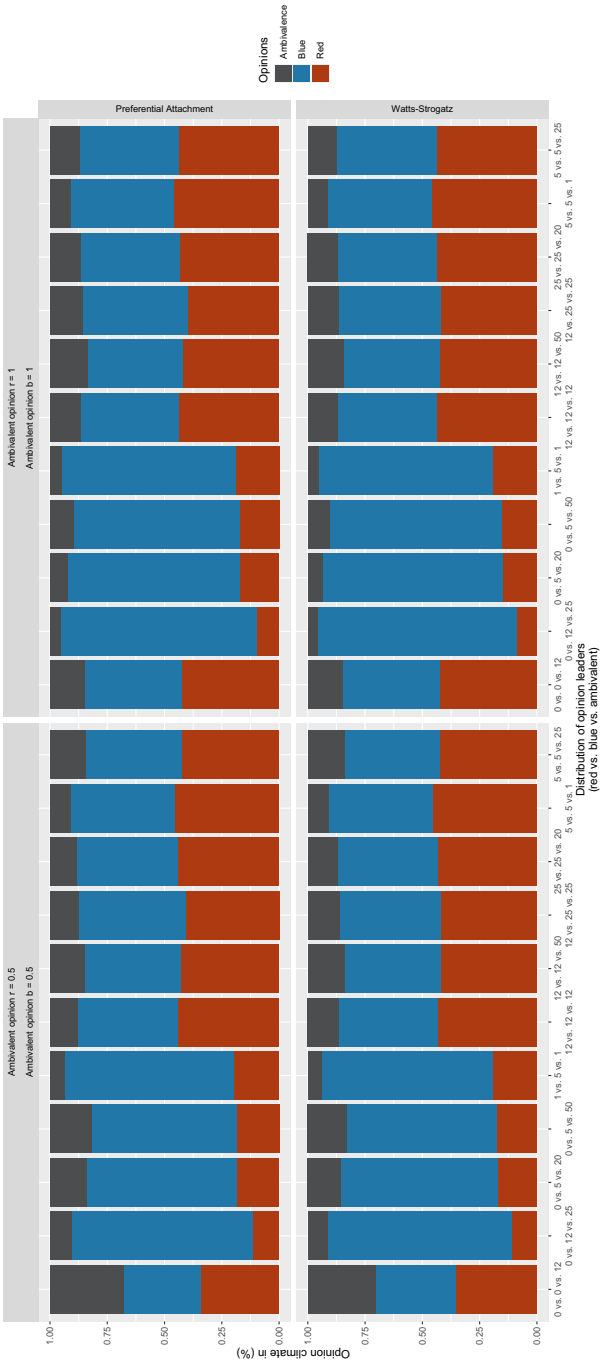

Fig. 8 Impact of ambivalent (left: moderate; right: strong) and strongly ambivalent opinion leaders as well as the interconnection with other opinion leaders

Table 5 Results of ambivalent opinion leaders interconnected with univalent opinion leaders

OL distribution	Network	Ambivalent value	% Blue	% Red	% Ambivalence
0 vs. 0 vs. 12	PA	$r=0.5, b=0.5$	33.84 CI [33.61, 34.08]	33.82 CI [33.58, 34.05]	32.34 CI [32.16, 32.52]
	WS	$r=0.5, b=0.5$	35.31 CI [35.14, 35.47]	35.11 CI [34.94, 35.28]	29.58 CI [29.46, 29.71]
	PA	$r=1, b=1$	42.35 CI [42.03, 42.68]	42.04 CI [41.72, 42.36]	15.61 CI [15.47, 15.75]
	WS	$r=1, b=1$	42.21 CI [42.02, 42.40]	42.24 CI [42.05, 42.43]	15.55 CI [15.45, 15.64]
0 vs. 5 vs. 20	PA	$r=0.5, b=0.5$	65.5 CI [65.34, 65.66]	18.25 CI [18.13, 18.38]	16.25 CI [16.16, 16.34]
	WS	$r=0.5, b=0.5$	68.64 CI [68.52, 68.76]	16.82 CI [16.72, 16.91]	14.54 CI [14.47, 14.62]
	PA	$r=1, b=1$	75.45 CI [75.29, 75.61]	16.82 CI [16.68, 16.96]	7.73 CI [7.67, 7.79]
	WS	$r=1, b=1$	78.59 CI [78.47, 78.72]	14.49 CI [14.38, 14.6]	6.92 CI [6.87, 6.97]
0 vs. 5 vs. 50	PA	$r=0.5, b=0.5$	62.98 CI [62.82, 63.14]	18.39 CI [18.26, 18.52]	18.63 CI [18.54, 18.72]
	WS	$r=0.5, b=0.5$	65.78 CI [65.65, 65.90]	17.06 CI [16.95, 17.16]	17.17 CI [17.09, 17.24]
	PA	$r=1, b=1$	72.39 CI [72.24, 72.54]	17.03 CI [16.89, 17.16]	10.58 CI [10.52, 10.64]
	WS	$r=1, b=1$	75.27 CI [75.15, 75.39]	14.86 CI [14.75, 14.96]	9.88 CI [9.83, 9.93]
0 vs. 12 vs. 25	PA	$r=0.5, b=0.5$	79.20 CI [79.08, 79.32]	11.29 CI [11.20, 11.38]	9.51 CI [9.44, 9.57]
	WS	$r=0.5, b=0.5$	80.01 CI [79.91, 80.11]	10.84 CI [10.76, 10.92]	9.15 CI [9.10, 9.20]
	PA	$r=1, b=1$	85.67 CI [85.57, 85.78]	9.51 CI [9.42, 9.59]	4.82 CI [4.78, 4.86]
	WS	$r=1, b=1$	86.9 CI [86.82, 86.99]	8.38 CI [8.31, 8.45]	4.72 CI [4.69, 4.75]
1 vs. 5 vs. 1	PA	$r=0.5, b=0.5$	73.8 CI [73.63, 73.96]	19.41 CI [19.27, 19.55]	6.79 CI [6.71., 6.87]
	WS	$r=0.5, b=0.5$	74.51 CI [74.40, 74.63]	19.1 CI [18.99, 19.2]	6.39 CI [6.33, 6.44]
	PA	$r=1, b=1$	75.75 CI [75.58, 75.92]	18.84 CI [18.70, 18.98]	5.41 CI [5.33, 5.48]
	WS	$r=1, b=1$	75.98 CI [75.86, 76.1]	18.83 CI [18.72, 18.94]	5.19 CI [5.15, 5.24]
5 vs. 5 vs. 1	PA	$r=0.5, b=0.5$	45.31 CI [45.19, 45.44]	45.29 CI [45.17, 45.42]	9.39 CI [9.32, 9.47]
	WS	$r=0.5, b=0.5$	45.13 CI [45.02, 45.24]	45.3 CI [45.19, 45.41]	9.57 CI [9.51, 9.64]
	PA	$r=1, b=1$	45.29 CI [45.16, 45.42]	45.53 CI [45.41, 45.66]	9.18 CI [9.10, 9.25]
	WS	$r=1, b=1$	45.32 CI [45.20, 45.43]	45.65 CI [45.54, 45.77]	9.03 CI [8.97, 9.09]

Table 5 (continued)

OL distribution	Network	Ambivalent value	% Blue	% Red	% Ambivalence
5 vs. 5 vs. 25	PA	$r=0.5, b=0.5$	42.08 CI [41.97, 42.19]	42.14 CI [42.03, 42.25]	15.78 CI [15.7, 15.86]
	WS	$r=0.5, b=0.5$	41.82 CI [41.71, 41.93]	42.04 CI [41.93, 42.15]	16.14 CI [16.07, 16.21]
	PA	$r=1, b=1$	43.34 CI [43.23, 43.46]	43.50 CI [43.38, 43.61]	13.16 CI [13.08, 13.24]
	WS	$r=1, b=1$	43.41 CI [43.30, 43.51]	43.6 CI [43.49, 43.7]	13 CI [12.93, 13.06]
12 vs. 12 vs. 12	PA	$r=0.5, b=0.5$	43.74 CI [43.63, 43.85]	43.90 CI [43.79, 44.02]	12.36 CI [12.29, 12.43]
	WS	$r=0.5, b=0.5$	43.21 CI [43.11, 43.30]	43.30 CI [43.08, 43.27]	13.62 CI [13.55, 13.68]
	PA	$r=1, b=1$	43.19 CI [43.08, 43.29]	43.30 CI [43.20, 43.41]	13.51 CI [13.44, 13.58]
	WS	$r=1, b=1$	43.51 CI [43.42, 43.6]	43.41 CI [43.32, 43.5]	13.08 CI [13.02, 13.14]
12 vs. 12 vs. 50	PA	$r=0.5, b=0.5$	42.32 CI [42.22, 42.43]	42.45 CI [42.35, 42.56]	15.22 CI [15.16, 15.29]
	WS	$r=0.5, b=0.5$	41.76 CI [41.66, 41.86]	41.82 CI [41.72, 41.92]	16.42 CI [16.36, 16.49]
	PA	$r=1, b=1$	41.64 CI [41.53, 41.74]	41.76 CI [41.66, 41.86]	16.61 CI [16.54, 16.67]
	WS	$r=1, b=1$	41.93 CI [41.84, 42.03]	41.94 CI [41.84, 42.03]	16.13 CI [16.07, 16.19]
12 vs. 25 vs. 25	PA	$r=0.5, b=0.5$	46.57 CI [46.45, 46.69]	40.61 CI [40.49, 40.72]	12.83 CI [12.77, 12.89]
	WS	$r=0.5, b=0.5$	44.16 CI [44.06, 44.26]	41.58 CI [41.48, 41.68]	14.26 CI [14.20, 14.32]
	PA	$r=1, b=1$	46.12 CI [46.01, 46.24]	39.5 CI [39.39, 39.61]	14.38 CI [14.31, 14.44]
	WS	$r=1, b=1$	44.53 CI [44.43, 44.62]	41.57 CI [41.48, 41.67]	13.9 CI [13.83, 13.96]
25 vs. 25 vs. 20	PA	$r=0.5, b=0.5$	43.92 CI [43.79, 44.05]	44.07 CI [43.93, 44.20]	12.01 CI [11.95, 12.07]
	WS	$r=0.5, b=0.5$	43.29 CI [43.20, 43.39]	43.18 CI [43.08, 43.27]	13.53 CI [13.47, 13.59]
	PA	$r=1, b=1$	43.13 CI [42.99, 43.26]	43.17 CI [43.03, 43.30]	13.71 CI [13.64, 13.77]
	WS	$r=1, b=1$	43.36 CI [43.26, 43.46]	43.43 CI [43.34, 43.53]	13.21 CI [13.14, 13.27]

PA, preferential attachment; *WS*, Watts–Strogatz

Compared to baseline (discreditation value: 0), where no discrediting takes place and the climate of opinion is symmetrical, the blue and red camps have across all distributions value of 49.26% CI [49.24%, 49.29%] for the preferential attachment and 49.46% CI [49.44%, 49.49%] for the Watts–Strogatz network. With an extreme discreditation value of –1, the size of the red opinion camp across all distributions has risen to 54.33% CI [54.23%, 54.43%] for the preferential attachment and 55.03%

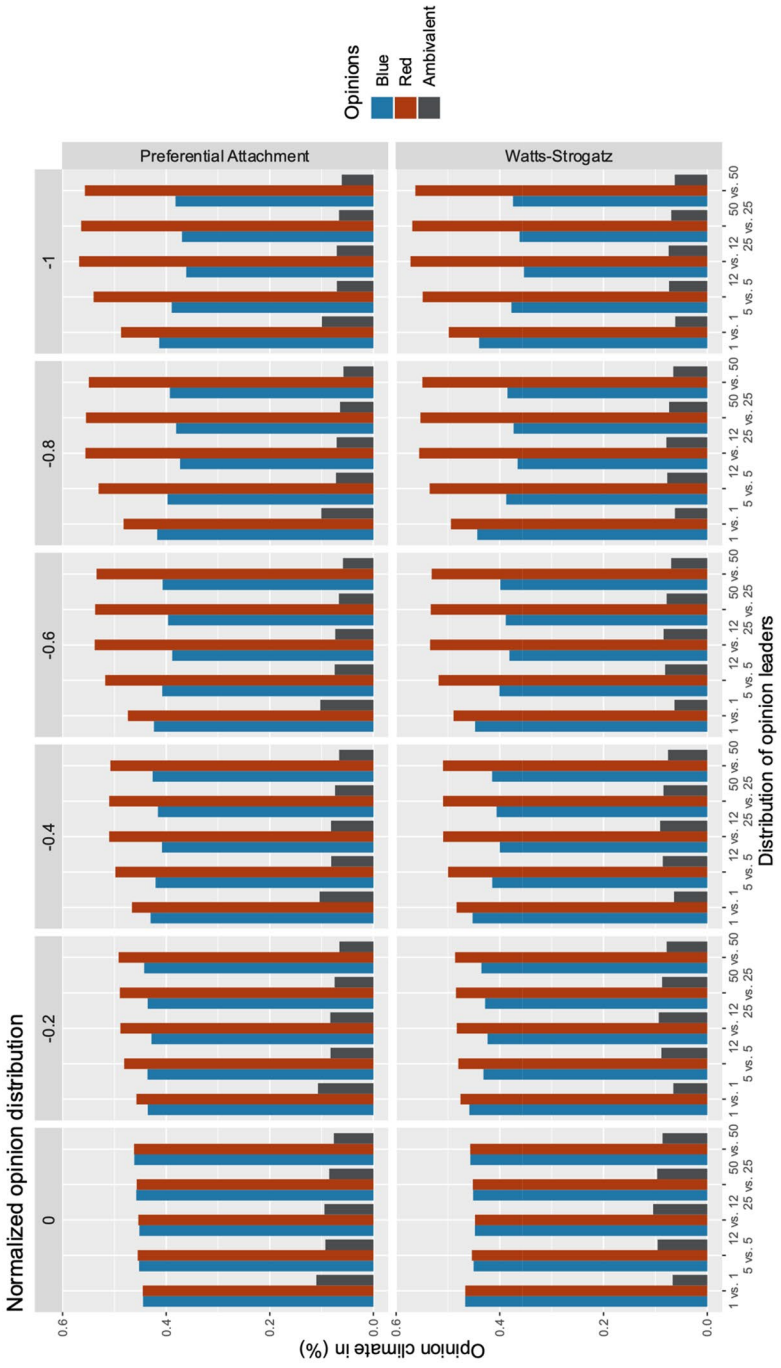

Fig. 9 The effect of discrediting opinion leaders on the distribution of opinion climate

CI [54.92%, 55.13%] for the Watts–Strogatz network, while the blue camp has decreased to 38.31% CI [38.23%, 38.39%] and 38.13% CI [38.05%, 38.22%]. Interestingly, there were also differences in the ratios of ambivalent agents when increasing the discrediting value. The percentage of ambivalent opinions decreased in the preferential attachment model from 9.18% CI [9.16%, 9.20%] to 7.36% CI [7.33%,

Table 6 Results of the modeling with different opinion leader (*OL*) distribution and network topologies to evaluate the opinion distribution

OL distribution	Network	Rabble value	% Blue	% Red	% Ambivalence
1 vs. 1	PA	$r=0$	44.46 CI [44.35, 44.58]	44.52 CI [44.4, 44.64]	11.02 CI [10.95, 11.09]
		$r=-0.2$	43.55 CI [43.43, 43.67]	45.74 CI [45.62, 45.87]	10.7 CI [10.63, 10.77]
		$r=-0.4$	43.01 CI [42.89, 43.13]	46.6 CI [46.48. 46.73]	10.38 CI [10.31, 10.45]
		$r=-0.6$	42.36 CI [42.24, 42.48]	47.39 CI [47.26, 47.53]	10.25 CI [10.18, 10.31]
		$r=-0.8$	41.73 CI [41.6, 41.86]	48.22 CI [48.08, 48.36]	10.05 CI [9.98, 10.12]
		$r=-1$	41.34 CI [41.21, 41.47]	48.71 CI [48.57, 48.86]	9.94 CI [9.87, 10.01]
	WS	r=0	46.67 CI [46.59, 46.74]	46.65 CI [46.58, 46.73]	6.68 CI [6.65, 6.71]
		$r=-0.2$	45.87 CI [45.8, 45.95]	47.58 CI [47.5, 47.65]	6.55 CI [6.51, 6–58]
		$r=-0.4$	45.25 CI [45.17, 45.33]	48.34 CI [48.25, 48.42]	6.41 CI [6.38, 6.44]
		$r=-0.6$	44.76 CI [44.67, 44.84]	48.92 CI [48.83, 49.01]	6.32 CI [6.29, 6.36]
		$r=-0.8$	44.34 CI [44.25, 44.43]	49.42 CI [49.33, 49.52]	6.24 CI [6.21, 6.27]
		r=−1	43.98 CI [43.89, 44.08]	49.84 CI [49.74, 49.94]	6.18 CI [6.14, 6.21]
5 vs. 5	PA	$r=0$	45.23 CI [45.18, 45.28]	45.5 CI [45.45, 45.55]	9.27 CI [9.24, 9.31]
		$r=-0.2$	43.61 CI [43.55, 43.67]	48.11 CI [48.04, 48.19]	8.28 CI [8.24, 8.32]
		$r=-0.4$	42.05 CI [41.96, 42.14]	49.81 CI [49.7, 49.92]	8.14 CI [8.1, 8.18]
		$r=-0.6$	40.76 CI [40.65, 40.88]	51.77 CI [51.62, 51.92]	7.47 CI [7.42, 7.52]
		$r=-0.8$	39.73 CI [39.59, 39.86]	53.05 CI [52.87, 53.23]	7.22 CI [7.17, 7.28]
		$r=-1$	38.94 CI [38.78, 39.09]	54.02 CI [53.82, 54.21]	7.05 CI [6.99, 7.11]

Table 6 (continued)

OL distribution	Network	Rabble value	% Blue	% Red	% Ambivalence
	WS	$r=0$	45.05 CI [45, 45.09]	45.38 CI [45.33, 45.42]	9.58 CI [9.54, 9.61]
		$r=-0.2$	43.15 CI [43.09, 43.21]	48 CI [47.92, 48.07]	8.85 CI [8.82, 8.89]
		$r=-0.4$	41.45 CI [41.36, 41.55]	49.97 CI [49.86, 50.09]	8.58 CI [8.54, 8.61]
		$r=-0.6$	40.07 CI [39.94, 40.19]	51.8 CI [51.62, 51.96]	8.13 CI [8.09, 8.17]
		$r=-0.8$	38.75 CI [38.6, 38.9]	53.53 CI [53.34, 53.72]	7.72 CI [7.67,7.77]
		$r=-1$	37.75 CI [37.58, 37.92]	54.88 CI [54.66, 55.1]	7.37 CI [7.32, 7.43]
12 vs. 12	PA	$r=0$	45.17 CI [45.13, 45.22]	45.38 CI [45.33,45.42]	9.45 CI [9.42, 9.48]
		$r=-0.2$	42.84 CI [42.77, 42.91]	48.84 CI [48.75, 48.93]	8.32 CI [8.29, 8.36]
		$r=-0.4$	40.81 CI [40.71, 40.92]	51.01 CI [50.87, 51.14]	8.18 CI [8.14, 8.22]
		$r=-0.6$	38.84 CI [38.69, 38.99]	53.8 CI [53.61, 53.99]	7.36 CI [7.31, 7.42]
		$r=-0.8$	37.32 CI [37.14, 37.5]	55.59 CI [55.36, 55.83]	7.09 CI [7.03, 7.15]
		$r=-1$	36.13 CI [35.92, 36.34]	56.8 CI [56.55, 57.06]	7.07 CI [7.01, 7.12]
	WS	$r=0$	44.79 CI [44.75, 44.84]	44.77 CI [44.73, 44.81]	10.43 CI [10.4, 10.46]
		$r=-0.2$	42.35 CI [42.29, 42.42]	48.3 CI [48.21,48.39]	9.34 CI [9.31, 9.38]
		$r=-0.4$	39.99 CI [39.88, 40.11]	50.92 CI [50.78, 51.06]	9.09 CI [9.05, 9.13]
		$r=-0.6$	38.13 CI [37.98, 38.28]	53.45 CI [53.25, 53.64]	8.42 CI [8.38, 8.47]
		$r=-0.8$	36.55 CI [36.37, 36.74]	55.56 CI [55.31, 55.8	7.89 CI [7.83, 7.95]
		$r=-1$	35.32 CI [35.11, 35.54]	57.23 CI [56.95, 57.51]	7.45 CI [7.38, 7.52]

Table 6 (continued)

OL distribution	Network	Rabble value	% Blue	% Red	% Ambivalence
25 vs. 25	PA	$r=0$	45.78 CI [45.72, 45.83]	45.69 CI [45.64, 45.75]	8.53 CI [8.51, 8.56]
		$r=-0.2$	43.57 CI [43.5, 43.65]	48.94 CI [48.85, 49.03]	7.49 CI [7.45, 7.52]
		$r=-0.4$	41.6 CI [41.49, 41.7]	51 CI [50.87, 51.13]	7.4 CI [7.37, 7.44]
		$r=-0.6$	39.63 CI [39.48, 39.77]	53.72 CI [53.54, 53.9]	6.65 CI [6.61, 6.7]
		$r=-0.8$	38.08 CI [37.9, 38.25]	55.49 CI [55.27, 55.71]	6.44 CI [6.39, 6.49]
		$r=-1$	36.95 CI [36.75, 3715]	56.41 CI [56.17, 56.66]	6.63 CI [6.59, 6.68]
	WS	$r=0$	45.14 CI [45.1, 45.18]	45.2 CI [45.16, 45.24]	9.66 CI [9.63, 9.69]
		$r=-0.2$	42.83 CI [42.76, 42.89]	48.45 CI [48.37, 48.53]	8.72 CI [8.68, 8.75]
		$r=-0.4$	40.61 CI [40.5, 40.72]	50.94 CI [50.81, 51.08]	8.45 CI [8.41, 8.48]
		$r=-0.6$	38.84 CI [38.69, 38.98]	53.32 CI [53.14, 53.51]	7.84 CI [7.79, 7.89]
		$r=-0.8$	37.33 CI [37.15, 37.51]	55.31 CI [55.08, 55.54]	7.36 CI [7.3, 7.41]
		$r=-1$	36.17 CI [35.97,36.38]	56.88 CI [56.62, 57.14]	6.94 CI [6.88, 7.01]
50 vs. 50	PA	$r=0$	46.16 CI [46.11, 46.21]	46.2 CI [46.15, 46.26]	7.64 CI [7.62, 7.66]
		$r=-0.2$	44.25 CI [44.18, 44.31]	49.19 CI [49.1, 49.27]	6.57 CI [6.53, 6.6]
		$r=-0.4$	42.62 CI [42.52, 42.71]	50.77 CI [50.65, 50.88]	6.62 CI [6.59, 6.65]
		$r=-0.6$	40.68 CI [40.55, 40.81]	53.44 CI [53.27,53.61]	5.88 CI [5.83, 5.92]
		$r=-0.8$	39.29 CI [39.13, 39.45]	54.93 CI [54.73, 54.13]	5.78 CI [5.73, 5.82]
		$r=-1$	38.2 CI [38.02, 38.38]	55.69 CI [55.48, 55.91]	6.1 CI [6.07, 6.14]

Table 6 (continued)

OL distribution	Network	Rabble value	% Blue	% Red	% Ambivalence
	WS	$r=0$	45.68 CI [45.64, 45.71]	45.7 CI [45.66, 45.74]	8.63 CI [8.6, 8.65]
		$r=-0.2$	43.54 CI [43.48, 43.6]	48.63 CI [48.55, 48.7]	7.84 CI [7.81, 7.86]
		$r=-0.4$	41.47 CI [41.37, 41.58]	50.96 CI [50.83, 51.08]	7.57 CI [7.54, 7.6]
		$r=-0.6$	39.89 CI [39.76, 40.03]	53.14 CI [52.98, 53.31]	6.96 CI [6.92, 7.01]
		$r=-0.8$	38.51 CI [38.35, 38.67]	54.93 CI [54.73, 55.14]	6.56 CI [6.51, 6.61]
		$r=-1$	37.44 CI [37.25, 37.63]	56.29 CI [56.06, 56.53]	6.27 CI [6.21, 6.32]

PA preferential attachment model, *WS* Watts–Strogatz, *CI* 95% confidence interval

7.39%] and in the Watts–Strogatz model from 8.99% CI [8.98%, 9.01%] to 6.84% CI [6.82%, 6.87%]. Thus, discrediting not only diminishes the discredited camp, but also the number of ambivalent nodes. To see a more detailed view about the different distributions, they can be seen in Table 6. Furthermore, the results have shown that the distribution of 12 vs. 12 OLs ensures that the opinion climate in the preferential attachment (51.90% CI [51.82%, 51.98%]), as well as in the Watts–Strogatz network (51.70% CI [51.62%, 51.79%]), develops most strongly in favor of the red camp across all negative discredit values.

Discussion

Applying agent-based modeling, this study investigated how OLs affect the opinion climate in online networks focusing on: (a) the impact of varying numbers of OLs in different opinion camps, (b) what influence ambivalent OLs exert in the network, and (c) how the opinion climate changes when OLs discredit the opposing opinion camp. Doing justice to the current state of knowledge in social psychology, these questions were examined in an opinion landscape in which individuals can have ambivalent opinions toward a certain issue. These scenarios are particularly relevant in the context of political communication, since OLs can spread their political stance (being univalent, ambivalent, or discrediting) in their network.

Addressing RQ1, we can state that an unequal distribution of OLs ensures that the opinion camp including the larger number of OLs—on average— "wins" over the opinion climate, i.e., the stance of the dominating OL is adopted by a majority of users in the network. Interestingly, however, even in the complete absence of OLs, the general opinion climate tips over time in favor of one side.

As our findings on RQ1 show, the numerically overrepresented group of OLs only dominates the general opinion climate when the other opinion camp is not represented by any OL. As soon as both reach a numerically critical mass (i.e., at least 12 OLs in each opinion camp), the advantage of the overrepresented camp disappears. This is in line with the previous findings that a few OLs can have a strong influence in the network and are responsible for the diffusion of opinions [21, 22, 95]. The finding can be explained by processes of complex contagion [96], i.e., the presence of multiple sources of influence within a complex social network structure. In such a structure, non-linear processes of influence often emerge. At the same time, the results show a "saturation effect," which reveals that at a certain point, an increasing number of OLs do not make a difference in how the opinion climate evolves. This observation may be explained by the idea of the "hard cores," that is, those who stick to their opinions in disregard of the external confirmations of other opinions, the conformity pressure exerted by others, and the potential social isolation that could be a consequence from being deviant from the majority [83, 97].

Our results also point to the boundaries of OL influence in complex and dynamically evolving social networks. These findings challenge the somewhat simplistic notion that OLs are primarily responsible for the formation of public opinion [1, 14], but rather put emphasis on the dynamically evolving social influence processes between “regular” users. In contrast to earlier findings (see [38]), our study shows that continuous increases in the number of OLs in a network only appear to impact public opinion under certain circumstances, i.e., when there is an extreme imbalance between OLs from opposing political camps. Furthermore, the impact of different shares of OLs who represent opposing stances appears to be dependent only to a limited extent on the network structure, with a somewhat higher amount of ambivalent users resulting in the preferential attachment network when no OLs are present at all (see [37]). Finally, network ambivalence (in terms of the share of attitudinally ambivalent users in the network) substantially decreases when there are only OLs advocating for one stance and remains largely unchanged when at least some OLs on each side are present. On one hand, this indicates that a network with an extreme imbalance of OLs from different opinion camps (e.g., within highly segregated networks) may foster further polarization over time [98], but on the other hand this shows that even a small number of opposing OLs may prevent a network from further segregation. These results are in line with the social psychological finding that consistently propagated minority views can have the potential to decrease majority influence [99] and emphasize that only some perseverant advocates of a minority stance can prevent a communication network from polarization.

In particular, when no opinion leader is represented, a difference between the two network topologies regarding the proportion of final ambivalent nodes was observable. A possible explanation of these results might be related to the strength of the weak ties in the different topologies [100]. Thus, it seems conceivable that the agents in the Watts–Strogatz network without opinion leaders and thus without a central hub are more weakly connected to each other in the network (weak ties), and thus, it is more difficult to come to a consensus of ambivalent opinions, since here the majority opinions of blue or red opinions have an advantage due to the ties. The preferential attachment model, on the other hand, has the characteristic that it has a

heterogeneous attachment and has already formed hubs, which allows strong ties to emerge, where ambivalent nodes in a continuous loop can also hear other continuous opinions, even if these are associated with majority opinions.

Regarding RQ2, it can be concluded that the greater the number of ambivalent OLs in the network, the more likely it is that opposing opinion camps will be equally represented. Previous research documented that when OLs hold a particular stance, they can affect the network in favor of their stance [101, 102]. However, as addressed in the present study, it seems conceivable that OLs can also represent two-sidedness, e.g., by arguing simultaneously in favor and against a certain issue. When this situation occurs, it is more likely that the overall opinion climate is more balanced and also that more individuals in the network hold ambivalent attitudes themselves. Interestingly, the presence of moderately ambivalent OLs appears to lead to somewhat higher levels of network ambivalence compared to highly ambivalent OLs. While this appears to be independent of the specific network topology, moderate strength of OLs' ambivalence may trigger just enough dynamic changes in the network to increase overall network ambivalence. Too many dynamic changes (caused by strongly ambivalent OLs), in contrast, lead to a somewhat lower network ambivalence. The degree of network ambivalence does not seem to increase with higher numbers of ambivalent OLs, but instead remains relatively stable when a certain threshold of OLs is reached. Again, this "saturation effect," here on the ambivalence level, might be explainable by the notion that "hard cores" might believe in the correctness of their opinion and are less susceptible to influence by other network members [83, 97] These findings indicate that social influence may not only be exerted by key network actors promoting one particular stance, but also by those who promote ambivalence and thereby increase the "integrative complexity" of their network [51]. Our simulation implies that political discussion networks (e.g., on social media) that are characterized by a polarized opinion climate (with most individuals clearly favoring one of two opposing issue stances or attitudes) may already benefit when only some influential actors (e.g., journalists and politicians) advocate for balanced views. However, while these actors may foster depolarization by increasing ambivalent attitudes in the whole network, the low magnitude of effects suggests that their influence is comparatively limited. From a normative perspective, this insight is particularly interesting, since it points to potential boundaries of the impact that unbiased or mainstream media and public players have on public opinion (see Lau et al., 2017 [103]).

With regard to RQ3, the results showed that OLs publishing discrediting messages are more likely to win over the opinion climate than if they solely advocated in favor of their stance (without discrediting the other side). The findings of this study also imply that the more vigorously an OL discredits or argues against the opposing side, the more likely they are to shape majorities in a network. These findings reflect the ideals proposed by public sphere theory [104, 105]: when individuals promote their own stance and debate by "debunking" the arguments of the other side, they may succeed in reaching a consensus within a network.

In terms of theoretical implications for OL research, the present study allows insights into the distribution of OLs and what effect ambivalent and discrediting OLs can have on the climate of opinion in the network. As Newman [106] points out,

understanding processes and behaviors in networks can help us understand complex phenomena that were previously difficult to explain. Our network analytic approach that modeled individual behavior as a function of external influence not only corroborates assumptions made by social psychology, predicting that one's social ties that hold divergent attitudes can increase one's personal attitudinal ambivalence [48, 50]. It also uncovered under which circumstances ambivalence can become a significant share within a network itself. While the ratio of opinion leaders and network topologies seem to be characteristics that shape the diffusion of ambivalence in a network, we also observed emergent phenomena such as the saturation effect of opinion leaders, potentially indicating that some actors are not susceptible to influence no matter how many hubs are present in a network. Implementation on a network level allows a consideration that is similar to social platforms such as Facebook or YouTube and thus takes into account the dynamic processes of the two-dimensional opinions of individuals, OLs, and their roles to measure their influence. Research on OLs usually focuses on their identification in networks [88, 107] and their specific characteristics [108]. Our results are intended to represent a theoretical basis for future research—integrating real-world data—to examine the relativizing effect of expressing ambivalent attitudes on the evolution of opinion climates. In particular, this research could serve to develop hypotheses in the field of political communication (e.g., in the context of election campaigns).

In terms of practical implications, results could play an important role in political and economic spheres. Political debates in online environments often appear to be clearly polarized and one-sided. Following the ideals of public sphere theory, it seems advisable to deal with the complexity, that is, the ambivalence of political attitudes and related arguments. Our research addresses this scenario and applies it to the existence of ambivalent OLs: based on our results, it appears that these ambivalent OLs can provide regulation of the opinion climate even when the proportions of OLs from two camps are different.

The findings regarding discrediting opinion leaders also reveal a practical implication in relation to contemporary journalism and news dissemination of information. The COVID-19 pandemic has once again highlighted the problematic presence of misinformation and its rapid spread on social media. Although the news reports on the scientific findings of the Coronavirus and their successes in the fight against the virus, misinformation still manages to get through to a portion of the population. Through social networks, discrediting opinion leaders who, for example, claim that the virus is harmless or that the government wants to control us with vaccination, can have an impact on their followers. To prevent this misinformation from influencing a wide proportion of the population, the government could implement countermeasures [109]. For example, journalists, public news houses, and influencers could serve as opinion leaders with a certain reach in their network and discredit counterarguments in order to fight misinformation.

Furthermore, the results can be used for specific business cases in the field of influencer marketing to promote product placements of opinion leaders more effectively and thus to identify and forecast the spread of opinions and future climates of opinion. This would allow companies to select which influencers are better suited

for product placement to have an advantage over competing products and to promote them to the most appropriate networks.

In terms of implications for future research, our model can be extended to include other factors to analyze other theories besides OLs such as the spiral of silence theory, in which a person's opinion expression behavior is influenced by their environment. The integration of OLs in this model could also shed light on the point at which certain individuals no longer influence the opinions of others. Our work extends previous research by focusing on opinion leadership in networks by representing opinions two-dimensionally and, in particular, analyzing the functionality of ambivalent and discrediting OLs. As a key point for further research, we believe that our model can be carried out with real network data from social platforms to validate the results of our simulations [34].

Limitations and further work

Our agent-based model solely represented two different camps of opinion, which can be compared, for example, to the political system in the United States, where Democrats and Republicans made up the lion's share of the opinion landscape. However, there are other political systems in which multiple parties exist, and therefore, more than two camps could be represented. Thus, future research could extend the existing binary modeling of opinions to include multiple opinions to study a more complex climate of opinion. In addition to the two applied network topologies, Watts–Strogatz and preferential attachment, other topologies could also be considered to make more specific statements on the role of network structure. In this context, we think that the investigation of the stochastic block models, which have the ability to form communities in graphs and to explicitly define their symmetric density, is particularly important. Using this model, findings showed that a low density in communities can lead to a reduction in the diversity of opinions [110]. In addition to the aspect of network topology, consideration could also be given to different centrality calculations, comparing the characteristics of OLs and the extent to which these change the climate of opinion. Besides the classic methods (closeness, betweenness) for identifying OLs, methods from game theoretical approaches might also be used.

Furthermore, future research could simulate the dynamics of opinion climates and the influence of OLs using real network data from platforms such as Facebook, YouTube, and Twitter. Here, a hybrid approach of sentiment analysis, network analysis, and agent-based modeling could be aimed at, in which a realistic representation of the opinion climate on different topics might be presented. One approach might be initially determining the opinions of users by means of an automatic sentiment analysis and filtering the individual topics on the basis of the thematic focuses. Next, the communication patterns (interactions) of the users with their calculated sentiment scores could be transformed into a network, which then serves as the basis for agent-based modeling. This could prove problematic in some cases, since the real data do not always fit 100% to the conditions of the network or vice versa. For example, although network data from social platforms show the flow of communication, a manual verification by humans would be required to ensure that opinions

are exchanged on topics and establish whether they are for or against something. Likewise, the input to the models may require parameters that are not reflected in the real data, such as OLs who have been shown to influence individuals in the network. Finally, in our modeling, specific numbers of opinion leaders (0, 12, and 25) were included. Even though these were based on empirical knowledge of the prevalence of OLs on social networking sites, future research could make use of a more fine-grained distinction with smaller increments in the number of OLs.

Conclusion

This study developed an agent-based model that deals with the phenomenon of opinion leadership and takes into account the ambivalence of different opinion camps on the basis of pertinent theoretical and empirical knowledge. The findings of this study show that OLs have an influence on the opinion climate, but that in particular, an extremely unequal distribution of OLs from different opinion camps leads to major adjustments (toward the dominating OL fraction) in the distribution of opinion climate on the network level. However, there appears to be a threshold value at which the imbalance of opposing OLs in the network no longer has any effect on the climate of opinion. This may indicate that in networks, social influence on the level of political opinions is subject to non-linear processes, where the global distribution of opinions changes in response to a critical number of opinion leaders from a political camp. To our knowledge, this is the first study that simulates processes of public deliberation and includes not only advocates of exclusively one viewpoint but also holders of ambivalent attitudes. We provide evidence that influential ambivalent players can increase ambivalence in individual users and counter processes of opinion polarization to a limited extent. Finally, we show that, when opinion leaders not only advocate for one side but also against the other side, the former opinion camp becomes more dominant within the network. Further research should model multiple attitude objects (i.e., more than one political issue) and systematically take the network structure and its dynamic changes into account (e.g., with regard to evolving clusters of like-minded users).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42001-022-00161-z.

Acknowledgements This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia (Grant Number: 005-1709-0004), Junior Research Group "Digital Citizenship in Network Technologies" (Project Number: 1706dgn009).

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. E. Katz und P. F. Lazarsfeld, *Personal Influence, The part played by people in the flow of mass communications*. Transaction publishers, 1966.
2. B. E. Weeks, A. Ardèvol-Abreu, und H. Gil de Zúñiga, „Online Influence? Social Media Use, Opinion Leadership, and Political Persuasion", *Int J Public Opin Res*, S. edv050, Dez. 2015, https://doi.org/10.1093/ijpor/edv050.
3. I. K. Schneider und N. Schwarz, „Mixed feelings: the case of ambivalence", *Current Opinion in Behavioral Sciences*, Bd. 15, S. 39–45, Juni 2017, https://doi.org/10.1016/j.cobeha.2017.05.012.
4. Nir, L. (2005). Ambivalent social networks and their consequences for participation. *International Journal of Public Opinion Research, 17*(4), 422–442. https://doi.org/10.1093/ijpor/edh069
5. C. J. Armitage, „Beyond attitudinal ambivalence: effects of belief homogeneity on attitude-intention-behaviour relations", *Eur. J. Soc. Psychol.*, Bd. 33, Nr. 4, S. 551–563, Juli 2003, https://doi.org/10.1002/ejsp.164.
6. Armitage, C. J., & Conner, M. (2000). Attitudinal ambivalence: a test of three key hypotheses. *Personality and Social Psychology Bulletin, 26*(11), 1421–1432. https://doi.org/10.1177/0146167200263009
7. Conner, M., & Sparks, P. (2002). Ambivalence and attitudes. *European Review of Social Psychology, 12*(1), 37–70. https://doi.org/10.1080/14792772143000012
8. Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus facebook: comparing incivility, impoliteness, and deliberative attributes. *New Media & Society, 20*(9), 3400–3419. https://doi.org/10.1177/1461444817749516
9. N. Ernst, S. Blassnig, S. Engesser, F. Büchel, und F. Esser, „Populists Prefer Social Media Over Talk Shows: An Analysis of Populist Messages and Stylistic Elements Across Six Countries", *Social Media + Society*, Bd. 5, Nr. 1, S. 205630511882335, Jan. 2019, https://doi.org/10.1177/2056305118823358.
10. D. Sohn, „Spiral of Silence in the Social Media Era: A Simulation Approach to the Interplay Between Social Networks and Mass Media", *Communication Research*, S. 009365021985651, Juni 2019, https://doi.org/10.1177/0093650219856510.
11. D. Sohn und N. Geidner, „Collective Dynamics of the Spiral of Silence: The Role of Ego-Network Size", *Int J Public Opin Res*, Bd. 28, Nr. 1, S. 25–45, März 2016, https://doi.org/10.1093/ijpor/edv005.
12. E. Bonabeau, „Agent-based modeling: Methods and techniques for simulating human systems", *Proceedings of the National Academy of Sciences*, Bd. 99, Nr. Supplement 3, S. 7280–7287, Mai 2002, https://doi.org/10.1073/pnas.082080899.
13. C. A. Bail *u. a.*, „Exposure to opposing views on social media can increase political polarization", *Proc Natl Acad Sci USA*, Bd. 115, Nr. 37, S. 9216–9221, Sep. 2018, https://doi.org/10.1073/pnas.1804840115.
14. I. Himelboim, E. Gleave, und M. Smith, „Discussion catalysts in online political discussions: Content importers and conversation starters", *Journal of Computer-Mediated Communication*, Bd. 14, Nr. 4, S. 771–789, Juli 2009, https://doi.org/10.1111/j.1083-6101.2009.01470.x.
15. R. Xiao, T. Yu, und J. Hou, „Modeling and Simulation of Opinion Natural Reversal Dynamics with Opinion Leader Based on HK Bounded Confidence Model", *Complexity*, Bd. 2020, S. 1–20, März 2020, https://doi.org/10.1155/2020/7360302.

16. P. F. Lazarsfeld, B. Berelson, und H. Gaudet, *The people's choice*. Oxford, England: Duell, Sloan & Pearce, 1944, S. vii, 178.
17. E. Katz, „The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis", *Public Opinion Quarterly*, Bd. 21, Nr. 1, Anniversary Issue Devoted to Twenty Years of Public Opinion Research, S. 61, 1957, https://doi.org/10.1086/266687.
18. R. M. Bond *u. a.*, „A 61-million-person experiment in social influence and political mobilization", *Nature*, Bd. 489, Nr. 7415, S. 295–298, Sep. 2012, https://doi.org/10.1038/nature11421.
19. Bode, L. (2016). Political news in the news feed: learning politics from social media. *Mass Communication and Society, 19*(1), 24–48. https://doi.org/10.1080/15205436.2015.1045149
20. Walter, S., & Brüggemann, M. (2018). Opportunity makes opinion leaders: analyzing the role of first-hand information in opinion leadership in social media networks. *Information, Communication & Society, 23*(2), 267–287. https://doi.org/10.1080/1369118X.2018.1500622
21. Valente, T. W., & Davis, R. L. (1999). Accelerating the diffusion of innovations using opinion leaders. *The ANNALS of the American Academy of Political and Social Science, 566*(1), 55–67. https://doi.org/10.1177/000271629956600105
22. Cho, Y., Hwang, J., & Lee, D. (2012). Identification of effective opinion leaders in the diffusion of technological innovation: a social network approach. *Technological Forecasting and Social Change, 79*(1), 97–106. https://doi.org/10.1016/j.techfore.2011.06.003
23. S. Choi, „The Two-Step Flow of Communication in Twitter-Based Public Forums", *Social Science Computer Review*, Bd. 33, Nr. 6, S. 696–711, Dez. 2015, https://doi.org/10.1177/0894439314556599.
24. King, C. W., & Summers, J. O. (1970). Overlap of opinion leadership across consumer product categories. *Journal of Marketing Research, 7*(1), 43–50.
25. P. S. van Eck, W. Jager, und P. S. H. Leeflang, „Opinion leaders' role in innovation diffusion: a simulation study: opinion leaders' role in innovation diffusion", *Journal of Product Innovation Management*, Bd. 28, Nr. 2, S. 187–203, März 2011, doi: https://doi.org/10.1111/j.1540-5885.2011.00791.x.
26. E. Bakshy, J. M. Hofman, W. A. Mason, und D. J. Watts, „Everyone's an influencer: quantifying influence on twitter", in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, S. 65–74.
27. Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology, 51*(3), 629–636. https://doi.org/10.1037/h0046408
28. B. Latan?, „The psychology of social impact.", *American Psychologist*, Bd. 36, Nr. 4, S. 343–356, 1981, https://doi.org/10.1037/0003-066X.36.4.343.
29. N. L. Abrica-Jacinto, E. Kurmyshev, und H. A. Juárez, „Effects of the Interaction Between Ideological Affinity and Psychological Reaction of Agents on the Opinion Dynamics in a Relative Agreement Model", *JASSS*, Bd. 20, Nr. 3, S. 3, 2017, https://doi.org/10.18564/jasss.3377.
30. G. Mckeown und N. Sheehy, „Mass Media and Polarisation Processes in the Bounded Confidence Model of Opinion Dynamics", *Journal of Artificial Societies and Social Simulation*, Bd. 9, Nr. 1, S. 11, 2006.
31. Railsback, S. F. (2019). *Agent-based and individual-based modeling: a practical introduction* (2nd ed.). Princeton University Press.
32. A. Waldherr und M. Wettstein, „Computational Communication Sciencel Bridging the Gaps: Using Agent-Based Modeling to Reconcile Data and Theory in Computational Communication Science", *International Journal of Communication*, Bd. 13, Nr. 0, 2019, [Online]. Verfügbar unter: https://ijoc.org/index.php/ijoc/article/view/10588
33. U. Wilensky und W. Rand, *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo*. Mit Press, 2015.
34. Alvarez-Galvez, J. (2016). Network models of minority opinion spreading: Using agent-based modeling to study possible scenarios of social contagion. *Social Science Computer Review, 34*(5), 567–581.
35. B. Scheufele, „Das Erklärungsdilemma der Medienwirkungsforschung.: Eine Logik zur theoretischen und methodischen Modellierung von Medienwirkungen auf die Meso- und Makro-Ebene", *Pub*, Bd. 53, Nr. 3, S. 339–361, Okt. 2008, https://doi.org/10.1007/PL00022227.
36. Squazzoni, F., Jager, W., & Edmonds, B. (2014). Social simulation in the social sciences: a brief overview. *Social Science Computer Review, 32*(3), 279–294.

37. Borowski, E., Chen, Y., & Mahmassani, H. (2020). Social media effects on sustainable mobility opinion diffusion: Model framework and implications for behavior change. *Travel Behaviour and Society, 19*, 170–183. https://doi.org/10.1016/j.tbs.2020.01.003
38. C. Kaiser, J. Kröckel, und F. Bodendorf, „Simulating the spread of opinions in online social networks when targeting opinion leaders“, *Inf Syst E-Bus Manage*, Bd. 11, Nr. 4, S. 597–621, Dez. 2013, https://doi.org/10.1007/s10257-012-0210-z.
39. Huckfeldt, R., Mendez, J. M., & Osborn, T. (2004). Disagreement, ambivalence, and engagement: the political consequences of heterogeneous networks. *Political Psychology, 25*(1), 65–95. https://doi.org/10.1111/j.1467-9221.2004.00357.x
40. Thompson, M. M., Zanna, M. P., & Griffin, D. W. (1995). Let's not be indifferent about (attitudinal) ambivalence. *Attitude strength: antecedents and consequences, 4*, 361–386.
41. K. Jonas, M. Diehl, und P. Brömer, „Effects of Attitudinal Ambivalence on Information Processing and Attitude-Intention Consistency“, *Journal of Experimental Social Psychology*, Bd. 33, Nr. 2, S. 190–210, März 1997, https://doi.org/10.1006/jesp.1996.1317.
42. I. R. Newby-Clark, I. McGregor, und M. P. Zanna, „Thinking and caring about cognitive inconsistency: When and for whom does attitudinal ambivalence feel uncomfortable?“, *Journal of personality and social psychology*, Bd. 82, Nr. 2, S. 157, 2002.
43. S. Feldman und J. Zaller, „The Political Culture of Ambivalence: Ideological Responses to the Welfare State“, *American Journal of Political Science*, Bd. 36, Nr. 1, S. 268, Feb. 1992, https://doi.org/10.2307/2111433.
44. Hohman, Z. P., Crano, W. D., Siegel, J. T., & Alvaro, E. M. (2014). Attitude ambivalence, friend norms, and adolescent drug use. *Prevention Science, 15*(1), 65–74. https://doi.org/10.1007/s11121-013-0368-8
45. D. C. Mutz, „The consequences of cross-cutting networks for political participation“, *American Journal of Political Science*, S. 838–855, 2002.
46. Dahlberg, L. (2004). The Habermasian public sphere: a specification of the idealized conditions of democratic communication. *Studies in social and political thought, 10*(10), 2–18.
47. P. E. Tetlock, „A value pluralism model of ideological reasoning.“, *Journal of personality and social psychology*, Bd. 50, Nr. 4, S. 819, 1986.
48. D. C. Mutz, *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press, 2006.
49. M. F. Meffert, M. Guge, und M. Lodge, „Good, bad, and ambivalent: The consequences of multidimensional political attitudes“, *Studies in public opinion: Attitudes, nonattitudes, measurement error, and change*, S. 63–92, 2004.
50. Petty, R. E., Tormala, Z. L., Briñol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from attitude change: an exploration of the PAST model. *Journal of Personality and Social Psychology, 90*(1), 21–41. https://doi.org/10.1037/0022-3514.90.1.21
51. Visser, P. S., & Mirabile, R. R. (2004). Attitudes in the social context: the impact of social network composition on individual-level attitude strength. *Journal of Personality and Social Psychology, 87*(6), 779–795. https://doi.org/10.1037/0022-3514.87.6.779
52. Levitan, L. C., & Visser, P. S. (2009). Social network composition and attitude strength: exploring the dynamics within newly formed social networks. *Journal of Experimental Social Psychology, 45*(5), 1057–1067. https://doi.org/10.1016/j.jesp.2009.06.001
53. S. S. Lee und B. K. Johnson, „Are they being authentic? The effects of self-disclosure and message sidedness on sponsored post effectiveness“, *International Journal of Advertising*, S. 1–24, Okt. 2021, https://doi.org/10.1080/02650487.2021.1986257.
54. U. Wilensky, *NetLogo. Evanston, IL: Center for connected learning and computer-based modeling, Northwestern University*. 1999. [Online]. Verfügbar unter: http://ccl.northwestern.edu/netlogo/
55. Thiele, J. C., Kurth, W., & Grimm, V. (2012). RNetLogo: an R package for running and exploring individual-based models implemented in NetLogo. *Methods in Ecology and Evolution, 3*(3), 480–483.
56. Clifford, P., & Sudbury, A. (1973). A model for spatial conflict. *Biometrika, 60*(3), 581–588. https://doi.org/10.1093/biomet/60.3.581
57. R. A. Holley und T. M. Liggett, „Ergodic theorems for weakly interacting infinite systems and the voter model“, *The annals of probability*, S. 643–663, 1975.
58. Sznajd-Weron, K., & Sznajd, J. (2000). Opinion evolution in closed community. *International Journal of Modern Physics C, 11*(06), 1157–1165.

59. DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association, 69*(345), 118–121.
60. G. Deffuant, D. Neau, F. Amblard, und G. Weisbuch, „Mixing beliefs among interacting agents“, *Advances in Complex Systems*, Bd. 3, Nr. 01n04, S. 87–98, 2000.
61. R. Hegselmann, U. Krause, und others, „Opinion dynamics and bounded confidence models, analysis, and simulation“, *Journal of artificial societies and social simulation*, Bd. 5, Nr. 3, 2002.
62. W. O. Kermack, A. G. McKendrick, und G. T. Walker, „A contribution to the mathematical theory of epidemics“, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, Bd. 115, Nr. 772, S. 700–721, 1927, https://doi.org/10.1098/rspa.1927.0118.
63. R. C. Tyson, S. D. Hamilton, A. S. Lo, B. O. Baumgaertner, und S. M. Krone, „The Timing and Nature of Behavioural Responses Affect the Course of an Epidemic“, *Bull Math Biol*, Bd. 82, Nr. 1, S. 14, Jan. 2020, https://doi.org/10.1007/s11538-019-00684-z.
64. Z. Wei und H. Ming-sheng, „Influence of opinion leaders on dynamics and diffusion of network public opinion“, in *2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings*, 2013, S. 139–144.
65. J. Woo, J. Son, und H. Chen, „An SIR model for violent topic diffusion in social media“, in *Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics*, 2011, S. 15–19.
66. J. Hou, T. Yu, und R. Xiao, „Structure Reversal of Online Public Opinion for the Heterogeneous Health Concerns under NIMBY Conflict Environmental Mass Events in China“, *Healthcare*, Bd. 8, Nr. 3, S. 324, Sep. 2020, https://doi.org/10.3390/healthcare8030324.
67. Yu, H., Cao, X., Liu, Z., & Li, Y. (2017). Identifying key nodes based on improved structural holes in complex networks. *Physica A: Statistical Mechanics and its Applications, 486*, 318–327. https://doi.org/10.1016/j.physa.2017.05.028
68. S. Schweighofer, D. Garcia, und F. Schweitzer, „An agent-based model of multi-dimensional opinion dynamics and opinion alignment“, *Chaos*, Bd. 30, Nr. 9, S. 093139, Sep. 2020, doi: https://doi.org/10.1063/5.0007523.
69. J. Li und R. Xiao, „Agent-Based Modelling Approach for Multidimensional Opinion Polarization in Collective Behaviour“, *JASSS*, Bd. 20, Nr. 2, S. 4, 2017, https://doi.org/10.18564/jasss.3385.
70. A.-L. Barabási und R. Albert, „Emergence of Scaling in Random Networks“, *Science*, Bd. 286, Nr. 5439, S. 509–512, Okt. 1999, https://doi.org/10.1126/science.286.5439.509.
71. D. J. Watts und S. H. Strogatz, „Collective dynamics of ‘small-world’ networks“, *Nature*, Bd. 393, Nr. 6684, S. 440–442, Juni 1998, https://doi.org/10.1038/30918.
72. Rahmandad, H., & Sterman, J. (2008). Heterogeneity and network structure in the dynamics of diffusion: comparing agent-based and differential equation models. *Management Science, 54*(5), 998–1014.
73. Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems, 28*(4), 394–412. https://doi.org/10.1080/0960085X.2018.1560920
74. Gandica, Y., del Castillo-Mussot, M., Vázquez, G. J., & Rojas, S. (2010). Continuous opinion model in small-world directed networks. *Physica A: Statistical Mechanics and its Applications, 389*(24), 5864–5870. https://doi.org/10.1016/j.physa.2010.08.025
75. A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, und B. Bhattacharjee, „Measurement and analysis of online social networks“, in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 2007, S. 29–42.
76. M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, und S. Moon, „I tube, you tube, everybody tubes: analyzing the world’s largest user generated content video system“, in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement - IMC ’07*, San Diego, California, USA, 2007, S. 1. https://doi.org/10.1145/1298306.1298309.
77. M. Wattenhofer, R. Wattenhofer, und Z. Zhu, „The YouTube Social Network“, *Proceedings of the International AAAI Conference on Web and Social Media*, Bd. 6, Nr. 1, Mai 2012, [Online]. Verfügbar unter: https://ojs.aaai.org/index.php/ICWSM/article/view/14243
78. A. Nazir, S. Raza, und C.-N. Chuah, „Unveiling facebook: a measurement study of social network based applications“, in *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement conference - IMC ’08*, Vouliagmeni, Greece, 2008, S. 43. https://doi.org/10.1145/1452520.1452527.

79. A. Vespignani, „Twenty years of network science“, *Nature*, Bd. 558, Nr. 7711, S. 528–529, Juni 2018, https://doi.org/10.1038/d41586-018-05444-y.
80. Hein, O., Schwind, M., & König, W. (2006). Scale-free networks: The impact of fat tailed degree distribution on diffusion and communication processes. *Wirtsch. Inform., 48*(4), 267–275. https://doi.org/10.1007/s11576-006-0058-2
81. L. Burbach, P. Halbach, M. Ziefle, und A. Calero Valdez, „Opinion Formation on the Internet: The Influence of Personality, Network Structure, and Content on Sharing Messages Online“, *Front. Artif. Intell.*, Bd. 3, S. 45, Juli 2020, https://doi.org/10.3389/frai.2020.00045.
82. F. Xiong, Y. Liu, und H.-F. Zhang, „Multi-source information diffusion in online social networks“, *Journal of Statistical Mechanics: Theory and Experiment*, Bd. 2015, Nr. 7, S. P07008, 2015.
83. Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of Communication, 24*(2), 43–51. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x
84. Miron, A. M., & Brehm, J. W. (2006). Reactance theory—40 years later. *Zeitschrift für Sozialpsychologie, 37*(1), 9–18. https://doi.org/10.1024/0044-3514.37.1.9
85. B. Nyhan und J. Reifler, „When Corrections Fail: The Persistence of Political Misperceptions“, *Polit Behav*, Bd. 32, Nr. 2, S. 303–330, Juni 2010, https://doi.org/10.1007/s11109-010-9112-2.
86. Risselada, H., Verhoef, P. C., & Bijmolt, T. H. A. (2016). Indicators of opinion leadership in customer networks: self-reports and degree centrality. *Marketing Letters, 27*(3), 449–460. https://doi.org/10.1007/s11002-015-9369-7
87. H. Zhang und X. Gong, „Leaders that bind: the role of network position and network density in opinion leaders' responsiveness to social influence“, *Asia Pacific Journal of Marketing and Logistics*, 2021.
88. F. Bodendorf und C. Kaiser, „Detecting Opinion Leaders and Trends in Online Communities“, in *2010 Fourth International Conference on Digital Society*, St. Maarten, Netherlands Antilles, Feb. 2010, S. 124–129. https://doi.org/10.1109/ICDS.2010.29.
89. W. Oueslati, S. Arrami, Z. Dhouioui, und M. Massaabi, „Opinion leaders' detection in dynamic social networks“, *Concurrency Computat Pract Exper*, Bd. 33, Nr. 1, Jan. 2021, https://doi.org/10.1002/cpe.5692.
90. A. Ligmann-Zielinska, D. B. Kramer, K. Spence Cheruvelil, und P. A. Soranno, „Using Uncertainty and Sensitivity Analyses in Socioecological Agent-Based Models to Improve Their Analytical Performance and Policy Relevance“, *PLoS ONE*, Bd. 9, Nr. 10, S. e109779, Okt. 2014, https://doi.org/10.1371/journal.pone.0109779.
91. J.-S. Lee *u. a.*, „The Complexities of Agent-Based Modeling Output Analysis“, *JASSS*, Bd. 18, Nr. 4, S. 4, 2015, https://doi.org/10.18564/jasss.2897.
92. Boccara, N. (2008). Models of opinion formation: influence of opinion leaders. *International Journal of Modern Physics C, 19*(01), 93–109.
93. X. Liu, C. Huang, H. Li, Q. Dai, und J. Yang, „The Combination of Pairwise and Group Interactions Promotes Consensus in Opinion Dynamics“, *Complexity*, Bd. 2021, 2021.
94. S. Wang und X. Fu, „Opinion dynamics on online-offline interacting networks: media influence and antagonistic interaction“, in *Proceedings of the 8th ACM International Workshop on Hot Topics in Planet-scale mObile computing and online Social neTworking - HotPOST '16*, Paderborn, Germany, 2016, S. 55–60. https://doi.org/10.1145/2944789.2944874.
95. D. Collings, A. A. Reeder, I. Adjali, P. Crocker, und M. H. Lyons, „Agent based customer modelling: individuals who learn from their environment“, in *Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No.00TH8512)*, 2000, Bd. 2, S. 1492–1497 Bd.2.
96. Centola, D., & Macy, M. (2007). Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology, 113*(3), 702–734. https://doi.org/10.1086/521848
97. J. Matthes, K. Rios Morrison, und C. Schemer, „A Spiral of Silence for Some: Attitude Certainty and the Expression of Political Minority Opinions“, *Communication Research*, Bd. 37, Nr. 6, S. 774–800, Dez. 2010, doi: https://doi.org/10.1177/0093650210362685.
98. Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.
99. Moscovici, S., Lage, E., & Naffrechoux, M. (1969). Influence of a consistent minority on the responses of a majority in a color perception task. *Sociometry, 32*(4), 365–380.
100. M. S. Granovetter, „The Strength of Weak Ties“, *American Journal of Sociology*, Bd. 78, Nr. 6, S. 1360–1380, Mai 1973, https://doi.org/10.1086/225469.

101. L. V. Casaló, C. Flavián, und S. Ibáñez-Sánchez, „Influencers on Instagram: Antecedents and consequences of opinion leadership“, *Journal of Business Research*, S. S0148296318303187, Juli 2018, https://doi.org/10.1016/j.jbusres.2018.07.005.
102. R. Huhn, J. Brantes Ferreira, A. Sabino de Freitas, und F. Leão, „The effects of social media opinion leaders' recommendations on followers' intention to buy“, *RBGN*, Bd. 20, Nr. 1, S. 57–73, Jan. 2018,https://doi.org/10.7819/rbgn.v20i1.3678.
103. R. R. Lau, D. J. Andersen, T. M. Ditonto, M. S. Kleinberg, und D. P. Redlawsk, „Effect of Media Environment Diversity and Advertising Tone on Information Search, Selective Exposure, and Affective Polarization“, *Polit Behav*, Bd. 39, Nr. 1, S. 231–255, März 2017, https://doi.org/10.1007/s11109-016-9354-8.
104. Dahlberg, L. (2005). The Habermasian public sphere: Taking difference seriously? *Theory and Society, 34*(2), 111–136.
105. Habermas, J. (1989). „The Structural Transformation of the Public Sphere: An inquiry into a category of bourgeois society, trans“, *Thomas Burger (Cambridge*. *Mass., 52*, 1989.
106. Newman, M. E. (2003). The structure and function of complex networks. *SIAM review, 45*(2), 167–256.
107. F. Li und T. C. Du, „Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs“, *Decision Support Systems*, Bd. 51, Nr. 1, Art. Nr. 1, Apr. 2011, https://doi.org/10.1016/j.dss.2010.12.007.
108. S. Winter und G. Neubaum, „Examining Characteristics of Opinion Leaders in Social Media: A Motivational Approach“, *Social Media + Society*, Bd. 2, Nr. 3, S. 205630511666585, Sep. 2016, https://doi.org/10.1177/2056305116665858.
109. S. Lewandowsky, U. K. H. Ecker, und J. Cook, „Beyond misinformation: Understanding and coping with the "post-truth" era.“, *Journal of Applied Research in Memory and Cognition*, Bd. 6, Nr. 4, S. 353–369, Dez. 2017, https://doi.org/10.1016/j.jarmac.2017.07.008.
110. S. Stern und G. Livan, „The Impact Of Noise And Topology On Opinion Dynamics In Social Networks“, *arXiv preprint arXiv:2010.12491*, 2020.

Research Paper 7: "The influence of community structure on opinion expression: an agent-based model"

Type	Journal
Rights and permission	Reproduced with permission from Springer Nature Open access
Authors	Cabrera, Benjamin; Ross, Björn; **Röchert, Daniel**; Brünker, Felix; Stieglitz, Stefan
Year	2021
Outlet	Journal of Business Economics
Publisher	Springer Nature
Permalink/DOI	https://doi.org/10.1007/s11573-021-01064-7
Full citation	Cabrera, B., Ross, B., Röchert, D., Brünker, F., & Stieglitz, S. (2021). The influence of community structure on opinion expression: an agent-based model. *Journal of Business Economics*, *91*(9), 1331-1355.

Journal of Business Economics (2021) 91:1331–1355
https://doi.org/10.1007/s11573-021-01064-7

ORIGINAL PAPER

The influence of community structure on opinion expression: an agent-based model

Benjamin Cabrera[1] · Björn Ross[1] · Daniel Röchert[1] · Felix Brünker[1] · Stefan Stieglitz[1]

Accepted: 8 September 2021 / Published online: 20 October 2021

Abstract
Social media has become important in shaping the public discourse on controversial topics. Many businesses therefore monitor different social media channels and try to react adequately to a potentially harmful opinion climate. Still, little is known about how opinions form in an increasingly connected world. The spiral of silence theory provides a way of explaining deviations between the perceived opinion climate and true beliefs of the public. However, the emergence of a spiral of silence on social media is hard to observe because only the thoughts of those who express their opinions are evident there. Recent research has therefore focused on modelling the processes behind the spiral of silence. A particular characteristic of social media networks is the presence of communities. Members of a community tend to be connected more with other members of the same community than with outsiders. Naturally, this might affect the development of public opinion. In the present article we investigate how the number of communities in a network and connectivity between them affects the perceived opinion climate. We find that higher connectivity between communities makes it more likely for a global spiral of silence to appear. Moreover, a network fragmented into more, smaller communities seems to provide more "safe spaces" for a minority opinion to prevail.

Keywords Agent-based model · Spiral of silence · Network · Communities · Stochastic block model

JEL Classification C15 · D85

✉ Benjamin Cabrera
benjamin.cabrera@uni-due.de

[1] University of Duisburg-Essen, Duisburg, Germany

1 Introduction

Businesses are influenced in many ways by the discussions that take place on social media. Sometimes, these take the form of short-lived social media storms that may threaten a company's reputation, for example in the case of a faulty product or claims of employee abuse, but they can also blow over quickly. At other times, these discussions can have serious long-term implications for individual businesses or for entire industries, with potential ripple effects throughout their supply chains. Current examples include debates on the carbon emissions of the various forms of private and public transport, such as cars and aeroplanes. These debates on social media do not only reflect the social norms around the usage of specific modes of transport, but they also shape them.

In order to make informed strategic decisions, it is essential for companies to understand the underlying processes shaping public opinion. One specific aspect of these processes is described by the spiral of silence theory (Noelle-Neumann 1974). It applies to controversial issues on which there are two opposing viewpoints, and it assumes that people's willingness to express their opinion (for example, for or against a new policy) depends on the opinions expressed by those around them. If they sense that their peers agree with them, they become more likely to voice their own opinion and the reverse. Over time, the theory posits, this behaviour can spiral into a situation in which a clear majority of publicly expressed opinions are in favour of one of the two viewpoints, and a consensus is established. A central point of the spiral of silence theory is that this apparent consensus opinion does not actually need to be held by the majority. The minority that espouses it might simply be more confident in their opinion and therefore more vocal.

This theory has been shown to apply to social media settings, to some extent. Recent simulation approaches show how individual decisions to spread one's opinion or not translate, on a macro level, into group dynamics that establish social norms (Ross et al. 2019). However, a key aspect of social media communication that has not been addressed in previous research, is that it often takes place in communities which are sometimes well-connected to other communities, but might also be more or less isolated. A real-life example of this would be a company that sells its products in various geographic markets, or to different segments of the population. This paper, therefore, explores how the development of a (perceived) public opinion, following the spiral of silence theory, is affected when the network is more or less subdivided into communities.

To form a better view of the spiral of silence process within and between online communities, we developed a new network model that is able to express various forms of community structure, and otherwise applied the same agent behaviour, regarding the spiral of silence, as in the work published by Ross et al. (2019). We use this model to investigate the influence of two key parameters: the size and number of communities (that is, whether the overall network is divided into many small ones or into few large ones), and the interconnectedness of the communities (that is, how many edges there are between communities, relative

to how many there could be). In network terminology, the first aspect relates to the distribution of nodes, the second to the distribution of edges. In the business context, the first question reflects how segmented the market is, while the second question reflects how close the markets or segments are to each other.

The paper makes the following contributions to the literature. Compared to previous studies on the spiral of silence, it clarifies how community structure affects this process. The more fragmented the network is into smaller communities, the less likely it is that one opinion is silenced entirely. The more interconnected the network, the more likely a global spiral of silence is to emerge in which only the supporters of one opinion are willing to express it. While previous studies on opinion dynamics in social networks that show community structure came to similar conclusions, these other studies were not based on the spiral of silence theory. Our work therefore shows that the assumptions of the spiral of silence can serve as an alternative explanation for this phenomenon. At the same time, it demonstrates the usefulness of agent-based modelling as a simulation technique.

The following section explains the necessary background information on how online opinion formation may impact business success, the role community structure plays in online social networks, and the spiral of silence theory. Section 3 describes the method, that is, the agent-based model and how it was validated. The results are shown in Sect. 4 and discussed in Sect. 5, with emphasis on their implications for research and for businesses.

2 Background

2.1 The impact of public opinion formation online on business success

In order to reduce uncertainty about their own decisions, customers often rely on information shared by others (Bikhchandani et al. 1992). Social media enhances the information exchange of product information and ratings by providing low-cost functionalities to reach a large audience and easily establish connections to other people. Thus, the formation of public opinion online has become even more important in the context of business success.

Research suggests that marketing managers observe public discourse to identify users' complaints and needs. This might lead to an improvement in the company's image. Likewise, the enhancement of a brand image could be due to the fact that the feelings and needs of customers are perceived and taken into account by the company's decision-makers (Kaiser et al. 2011). However, it is crucial to detect upcoming negative opinions towards a product, brand or person before they spread in the network or community. The detection and counteraction are particularly important due to the finding that customers' opinions could be influenced by the opinions of others (Sunder et al. 2019). In general, customers adapt their reviews or ratings of products to the general opinion of the crowd (Muchnik et al. 2013; Jiang and Wu 2017). Furthermore, if two reference groups have distinct opinions, experienced users rely more on their friends' opinions than on the crowd. In contrast, new or inexperienced users, who have had less time to establish strong connections to others, rely more on

the crowd than on new friends on social media (Sunder et al. 2019). In this context, people might weigh their connection to others such as the crowd, media outlets, or opinion leaders, that is, users who are likely to influence other users within their personal network (Jiang and Wu 2017; Watts and Dodds 2007; Katz and Lazarsfeld 1955).

2.2 Community structure in online social networks

The possibility to get in contact with people across temporal and spatial distances and to communicate with them is omnipresent. Social online platforms such as Twitter, Facebook or YouTube provide opportunities for networking and community formation. In this context, the term community describes a group of nodes (that is, users or accounts) that are more strongly connected to each other than to the rest of the network.

The investigation of political communities in terms of their participation has already been the goal of many studies (Grace-Farfaglia et al. 2006; Oser et al. 2013; Velasquez 2012). Similarly, studies found that national cultures also differ in their active participation in online political communities, according to the findings, 13.7% of Americans participate in political communities, compared to 7.45% in the Netherlands and 6.1% in South Korea (Grace-Farfaglia et al. 2006).

In addition, there are various reasons and motives for people to join communities, e.g., information and social friendship building (Ridings et al. 2006) or gaining a deeper understanding of the opinions and attitudes of others (Herring 1996). The exchange of political or ideological opinions, in which the individual's point of view is reinforced, can lead to the emergence of virtually homogeneous spaces, so-called "echo chambers", wherein like-minded people interact *only* with each other (Boutyline and Willer 2017). It is also argued that such self-reinforcing "echo chambers" can be seen as a danger to society, as they are particularly associated with polarisation and radicalisation because users are more extreme in their views (Prior 2007). Seen in this way, homophily, that is, "the principle that a contact between similar people occurs at a higher rate than among dissimilar people" (McPherson et al. 2001, p. 416), could also be a reason for the formation of communities.

Political or opinion-based homogeneous discussion areas have already been examined on Facebook (Bakshy et al. 2015), YouTube (Röchert et al. 2020) and Twitter (Barberá et al. 2015), but the aspect of the individual communities within the network has largely been ignored. Williams et al. (2015) analysed the Twitter communication network on climate change using a network analysis and found that there is a strong homogeneity in the interactions between like-minded communities of climate change activists and climate sceptics. More specifically, climate change activists expressed positive opinions with each other, while climate sceptics expressed negative opinions among themselves. Furthermore, the authors were able to identify mixed communities, which were characterized by a balanced and polarized content. The results of homogeneous communities are in line with the results of Conover et al. (2011), who found that the retweet network for political communication on

Twitter during the 2010 U.S. midterm elections was very polarized, with only few connections between left- and right-leaning users.

2.3 The spiral of silence

The spiral of silence theory (Noelle-Neumann 1974) explains changes in people's willingness to express their opinion as the result of a fear of being socially isolated. People sense the opinions on controversial topics of those around them and modify their public behaviour accordingly. Over time, this results in the formation of a consensus, the establishment of a social norm. Crucially, this consensus opinion does not even need to be held by the majority. It could simply be the case that the minority that holds this opinion is especially vocal about it, or is using especially effective communication channels to reach many people, leading the actual majority to not express their opinion openly. The assumptions of this theory have been the subject of much empirical research. In the realm of social media, it has been shown that individuals are (slightly) affected in their assessment of the overall opinion distribution by what they see online (Neubaum and Krämer 2017).

Although it has frequently been applied to attitudes towards political questions such as capital punishment, the spiral of silence theory in the original conceptualisation has always applied to a wide range of social norms including, for example, homeowners shovelling snow from their share of the sidewalk (Noelle-Neumann 1974). If they affect consumption decisions, the social pressures and group norms explained by this theory can have long-term strategic implications for businesses and entire industries.

A recent example is the Swedish concept of "flygskam" (flying shame or flight shame). Aware of the impact of air travel on carbon emissions, the environmentally conscious switch to other modes of transport (Weston et al. 2019). This choice is often communicated publicly, for example on social media. Climate activist Greta Thunberg's decision to sail to the UN climate summit by yacht instead of flying was widely and controversially discussed in traditional and social media (Parker, 2019). Although empirical evidence does not yet indicate that those with a higher awareness of climate change have lower greenhouse gas emissions from flights—if anything, the opposite is the case (Czepkiewicz et al. 2019)—and although its reach is geographically limited, with many markets for air travel, such as the Asia–Pacific Region, experiencing unprecedented growth (IATA 2018), if this movement grows it might threaten air travel as a leisure activity. According to news reports, airline executives were already worried in 2019 (Rucinski et al. 2019). In 2020–2021, the COVID-19 pandemic severely hit the airline industry. Although at the time of writing, its permanent effects are still hard to predict, it seems likely that some peer groups will exert additional social pressure against long-distance travel to avoid spreading the disease.

A related example is that of choosing to drive a car and choosing which car to drive. In a study by Hopkins (2016) in New Zealand, the Generation Y interviewees were highly aware of the environmental impact of cars, especially to commute, and

some participants decided not to drive for environmental reasons despite owning a driver's license.

A guilty conscience due to the perceived negative effects of transport choice would not in itself be enough to meet the theoretical assumptions of the spiral of silence. The theory does not predict changes in people's privately held opinions, but in the expressions of these opinions. The individuals in question would need to fear being socially isolated as a result of their choices. Indeed, previous research has at least surmised a link between social status and environmentally conscious consumption choices. Kahn (2007) showed a difference between environmentally indifferent "brown" communities and green ones where "the group norm is to live a sustainable lifestyle … driving a Prius would increase one's status while driving a Hummer would have the opposite effect". In making this link, we assume that the communication of one's opinion does not necessarily need to happen verbally, since the choice to buy and drive a car is an equally public display of one's attitudes.

2.4 Related work on simulating opinion dynamics

Complex communication processes such as the spiral of silence are challenging to investigate due to the complexity of empirical test procedures, as they require many resources, such as long-term observations of experiments and also the need for a large number of participants (Waldherr and Wettstein 2019). With the help of agent-based modelling, it is possible to investigate social phenomena of micro-level findings at the macro level (Epstein 2006; Klein et al. 2018), such as the dynamic processes within a network and how the interactions within the agents develop (Bruch and Atwell 2015).

In the literature on opinion dynamics, a variety of different methods based on mathematical and physical rules exist to simulate the mechanisms of interaction and their influence on opinions (Castellano et al. 2009). However, because there are many different modelling decisions to make, research questions to answer and results to focus on, a comprehensive overview of the field of opinion dynamics would be out of scope here. The following paragraphs, therefore, each focus on a different aspect of the research, namely the different types of interactions between opinions used in different models, the effect of communities and existing models studying the spiral of silence.

Models simulating opinion dynamics can broadly be split into those modelling opinions as discrete values [the Voter model (Clifford and Sudbury 1973; Holley and Liggett 1975), Snajzd model (Sznajd-Weron and Sznajd 2000)] and those using some form of continuous representation of opinions. The main difference between these two categories is the range of opinions that the agents in the models can assume. Discrete models often have binary opinions (i.e., in favour, against), while in continuous models, opinions are scalars, often bounded by an interval of values.

A well-studied class of continuous models are so-called bounded confidence models (Deffuant et al. 2000; Hegselmann and Krause 2002). These models consist of a set of agents, each of which is assigned an opinion, modelled as a real value in the interval [0,1]. The difference between these models lies in their view of how

they implement the communication between individuals. In the Deffuant model, the dynamic is based on the interaction of two individuals randomly connected to each other in the network, while the HK model considers the interactions of individuals in larger groups. The opinion values change over time depending on the values of the other agents, and the strength of the connections between the agents. The non-linearity and "boundedness" of the models is introduced by the fact that an agent only considers opinions that deviate up to a bound from its own opinion value. Interestingly, depending on the initial configuration and model parameters, the formation of clusters of agents with similar opinions can be observed (Lorenz 2006). Another classical continuous model, the DeGroot model, can be classified in the category of averaging models, in which agents determine their opinion based on the average of their neighbours' opinions (DeGroot 1974). This model has been used as the foundation for other models, such as the Friedkin-Johnsen model, which takes into account the aspect of stubbornness, so that individuals hold on to their original opinion to a certain level (Friedkin and Johnsen 1990). In a study in the area of interpersonal social influence, Ye et al. (2019) explicitly distinguish between private and expressed opinions in order to identify how they can deviate from another over time. Here, they used a strongly connected, aperiodic directed network to show that the combination of the network's strong interconnectedness, the individual's pressure to conform, and the individual's stubbornness have an impact on the discrepancy.

While the previously mentioned articles focus mainly on the different ways opinion interaction can be modelled, other research modelling opinion dynamics has specifically focused on studying the effects of community structure. The non-linear model of Banisch and Olbrich (2019), based on reinforcement learning (Q-learning), addresses the question of how bi-polarised opinion distributions can emerge and persist. In their model, the individual agents learn of and adapt to the opinions of their neighbours. To model the social structure, a random geometric graph was used in which agents communicate with those they are physically close to. The presence of communities and structural holes is a key aspect that allows the formation of a stable polarised opinion climate: in dense, less modular networks, polarisation disappears in favour of a global consensus. In a recent study by Stern and Livan (2021), the DeGroot and Friedkin-Johnsen models were used and extended to investigate the diversity of opinions in networks. Here, the network structure used was a stochastic block model; their results showed that the diversity of opinions decreases due to closed communities and thus it is more difficult to come to a common consensus.

There are already models specifically focussed on simulating the circumstances around the spiral of silence theory. In an article by Wu et al. (2015), an agent-based model is proposed where each agent is initialized holding one of two opinions. Then, a single agent is selected (the "first speaker") who expresses its opinion and triggers its neighbours to either also express their opinion, or stay silenced, depending on an "opinion pressure" which is based on the network topology around the agents and their neighbours opinions. Agents that have either expressed their opinion or stayed silent become "immune" and can't be triggered again. This process continues until no agents are left that could trigger a response. This model was then used to study the global opinion distribution based on different network topologies. The spiral of silence was also examined in a setting of more complex agent behaviour by

Sohn and Geidner (2016). Here, the agents randomly move around on a two-dimensional plane and express their opinion (if they are confident enough to do so) only to those agents that are physically close to them. In the model by Ross et al. (2019), inspired by the agent behaviour used by Sohn and Geidner (2016), the agents' influence is determined by a small-world, scale-free network that is more representative of online social networks and a world with internet-based communication. A recent study by Ma and Zhang (2021) used agent-based simulation for a model of opinion expression dynamics inspired by the spiral of silence theory in that people's willingness to express their opinion depends on perceived peer support for that opinion. Since their goal was to model a social media chat group where every user sees every other user's posts, their simulation assumed a fully connected network.

The spiral of silence networks studied by Sohn and Geidner (2016), Ross et al. (2019), and Ma and Zhang (2021) do not exhibit a community structure. However, real social networks exhibit varying amounts of modularity (Guerra et al. 2013), and a good simulation model for the discussed cases should therefore directly take into account community structure.

3 Methods

To test the effects of community structure on the spiral of silence process, we used an agent-based simulation model. Agent-based models are used in a variety of disciplines, from physics and biology to the social sciences (Wilensky and Rand 2015). Their strength lies in their versatility. The user specifies an environment and agents that populate it, including the rules according to which the individual agents act and react to their surroundings. It is then possible to observe the model as time passes, to pause and inspect the model and to calculate various statistics at any point in time.

This simulation approach has various advantages over other approaches that study social media usage. The simulation can be carried out an arbitrary number of times and populated with an arbitrary number of agents, unlike laboratory experiments, which are unfeasible for opinion formation processes in very large groups. Quasi-experimental studies in which the behaviour of many is manipulated are morally questionable (Flick 2016). Finally, unlike in large-scale observational studies, it is possible to examine the variables of agents, that is, to peer into the minds of those who remain silent. Otherwise, if messages such as Facebook posts and tweets are examined, the study would be limited to examining the opinions of those who are willing to express them. However, a critical challenge when working with agent-based models is to ensure that they accurately reflect reality. This section describes the network model, agent behaviour, and validation measures taken.

The simulation model is largely based on Ross et al. (2019)'s. The key difference is that the network model includes subcommunities. We also reimplemented the model in C++, as this allows for much faster run times and more flexibility compared to the original NetLogo implementation. The ability to perform more

simulation runs translates into an increased precision of results. The source code is freely available on GitHub https://github.com/bencabrera/spiral_of_silence_abm.

3.1 Modelling networks with cohesive communities

Networks are at the heart of our agent-based model for simulating the spiral of silence. They define the interaction topology of the agents by determining which agents another agent considers when gauging the opinion climate. Since the process of producing results from an agent-based model involves averaging outcomes of many simulations, a network model is needed to randomly generate new instances to run the simulations on.

This work differs from Ross et al. (2019) in the network model used. Ross et al. (2019) used a preferential attachment model (Albert and Barabási 2002), as it creates power-law tailed degree distributions typical for social networks (Barabási and Albert 1999) and is often considered a good, albeit simple, model of social networks in general (Newman 2003). The focus of this work, in contrast, lies on studying the effects of network communities on the spiral of silence. It is therefore essential to have a method of reliably generating networks containing ground-truth communities, which the simple preferential attachment models are not capable of.

The following paragraphs contain a brief review of important definitions. First of all, we only consider undirected networks, formally characterized by the mathematical notion of an undirected graph $G = (V, E)$, where V is a set of agents (or nodes) and $E \subset \{\{u, v\} : u, v \in V\}$ the set of connections (or edges). This implies that any influence between two agents runs both ways. If agent A and agent B are connected by an edge, then agent A is influencing B as well as the other way around. The *density* ρ of a network is the number of edges in the network divided by the number of possible edges, i.e. $\rho = \frac{2|E|}{|V|(|V|-1)}$.

A community structure in a network (Wasserman and Faust 1994) can be described by partitioning the nodes into multiple subsets $V_i \subset V$ (the communities) according to some characteristics. While Wasserman and Faust (1994) propose different such characteristics, the one most commonly used in empirical network analysis (Fortunato 2010) is based on the idea that network communities ought to have more connections between members of the same community (*intra-community edges*) than between members of different communities (*inter-community edges*). Analogue to the density, one can also define the *intra-community density* ρ_{in} and the *inter-community density* ρ_{out} as the ratios between the number of intra-/inter-community edges and the number of all possible edges of the respective type. With these definitions in place, we say a node partition actually represents a community structure if ρ_{in} is significantly higher than ρ_{out}, with the meaning of "significantly" depending on the actual case at hand. With the notion of network communities defined, it is still unclear how networks with such communities can be generated reliably.

The most common type of network models used to generate networks with communities is stochastic block models (SBM). Different variants of stochastic block models have been proposed (Abbe 2017). However, generally, they take at least three parameters: the total number of nodes n, a partition of the node set into r communities, and a $r \times r$ matrix P of probabilities, where $P_{i,j}$ specifies the probability of connecting nodes from community i with nodes from community j. Sometimes the partition into communities is also sampled from a given probability distribution (Abbe 2017). The network is then built by randomly deciding for every pair of vertices independently if the two vertices should be connected by an edge or not, using the P matrix entries as explained before. Note that the diagonal values $P_{i,i}$ characterise the probability of connecting two nodes inside the same community i, while the $P_{i,j}$ for $i \neq j$ are the probabilities of connecting nodes of different communities. This implies that the diagonal values of P are typically chosen much larger than the off-diagonal ones, in order to create cohesive communities. Also, for a constant probability matrix $P_{i,j} = c$ for all $1 <= i,j <= r$, the model is equal to an Erdös-Renyi model (Erdős and Rényi 1959) with parameter c, and therefore no community structure would be visible.

This classic Stochastic Block Model has several drawbacks that lead us to use a slightly different model for generating networks. First, networks generated from a classic SBM do not exhibit a power-law tailed degree distribution, typical for social networks. Instead, in the classic SBM every community is essentially an Erdös–Renyi random graph that has a Poisson distribution of the degrees (Newman 2003). An even bigger problem with a classic SBM is the fact that the generated networks are not necessarily connected. Especially for the targeted densities, Erdös-Renyi graphs tend to break down into many disconnected components, which is not desired in an agent-based model that relies on connections for propagating influence to other agents.

To solve these problems, we use a different model for generating networks with communities. Again, we take as parameters the partition of n nodes into r communities. However, the mechanisms for generating intra- and inter-community edges are now different. We generate every community based on the preferential attachment model by Barabási and Albert (1999), which takes a parameter m that defines the number of connections a new node makes to existing nodes when added to the network. The inter-community edges are then sampled similarly as before by randomly deciding for every pair of nodes in different communities if they should be connected by an edge or not. The probability of connecting two nodes of different communities is based on a third parameter ρ_{out}. It is no coincidence that this parameter is called ρ_{out}, as the inter-community density on a network generated by this model will on average be ρ_{out}. With the described model we will almost always generate a connected network (at least for reasonable values of ρ_{out}). Moreover, the network will have a power-law tailed degree distribution if the number of communities is significantly lower than the number of total nodes in the network. The parameter ρ_{out} has to be chosen with care in order to compare different networks with each other because the number of communities and their sizes affect the intra-community density and ρ_{out} has to be selected appropriately to match, for example, a targeted density ρ of the network

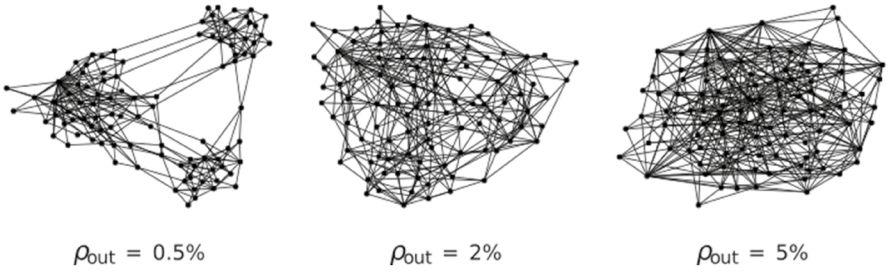

Fig. 1 Three networks generated with the stochastic block model used in the simulations. All networks have 100 nodes and an a-priori (50, 25, 25) partition into communities. Every community is generated by the Barabási-Albert model with $m = 3$. From left to right, higher inter-community densities ρ_{out} are used. In the left drawing, the three communities are clearly visible. The network in the middle already has many inter-community edges, such that the communities start to blend and in the rightmost drawing almost no communities are visible

as a whole. Figure 1 shows three networks with three communities each, differing only in the connectivity between communities.

3.2 Agent behaviour

While the previous section discussed the model of *which* agents are influencing which other agents, another important decision to make when creating an agent-based model is *how* the agents interact with each other. Agents in our model behave as they do in the work of Ross et al. (2019). The following section will, therefore, only briefly review the model and the empirical research its assumptions are based on.

The model simulates a simplified scenario of opinion formation. It is assumed that every agent i has an *opinion* o_i with values in $\{+, -\}$, representing either a positive or negative stance regarding a specific topic. Moreover, the opinion is fixed in time, that is, for the whole duration of the simulation an agent always has the same opinion. The opinion is randomly initialized as uniformly across the population, that is, every agent has equal probability of getting + or − as their opinion. This is because the model is not used to study how people change their opinions, but rather how people can become silenced when they feel that their opinion is not adequately represented in the population. The second property of an agent i is its *willingness to self-censor* Φ_i (Hayes et al. 2005, 2010). It determines whether an agent is easily silenced or holds their opinion even in the face of overwhelming opposition. It is also constant in time, as we assume this to be a relatively stable characteristic of a person. In our experiments, the willingness to self-censor is initialized for each agent as a uniformly distributed random value in [0,1].

The next property of an agent is its *confidence* $c_i(t)$. After each step, it is compared to the *willingness to self-censor* Φ_i and, if greater, the agent communicates its opinion to its surroundings (and is called *speaking*), whereas if smaller, the agent is *silenced* and does not speak out its opinion. It changes over the course of a simulation depending on the opinion climate surrounding an agent. Accordingly, the

opinion climate at time t, $\delta_i(t)$, observed by an agent i, is used to update its confidence. It is defined as $\delta_i(t) = \frac{n_s(i,t)-n_o(i,t)}{n_s(i,t)+n_o(i,t)}$, where $n_s(i,t)$ is the number of neighbours of agent i openly supporting its opinion, while $n_o(i,t)$ is the number of neighbours openly opposing it. There is no change to an agent's confidence ($\delta_i(t) = 0$) when its neighborhood is completely silent. Confidence is updated as follows: $c_i(t) = 2 \times \left(1 + e^{-\hat{c}_i(t)}\right)^{-1}$, where $\hat{c}_i(t) = max\left\{\hat{c}_i(t-1) + \delta_i(t); 0\right\}$. The value $\hat{c}_i(t)$ is initialized as a uniformly distributed random value in [0,1]. The transformation into $c_i(t)$ ensures that it stays within this range. As a result of these definitions, if there are more agents in the neighborhood of an agent i expressing their support for the opinion of i than there are agents opposing it, then δ_i will be positive and agent i's confidence increases. Similarly, δ_i is negative if there is more opposition than support in the neighbourhood of agent i and its confidence will drop. It should also be emphasised that only non-silenced neighbours are considered when computing $\delta_i(t)$. Silenced agents have no influence on the opinion climate, which is also why an agent becoming silenced can trigger a cascade of multiple agents becoming silenced or speaking again. In line with previous empirical research, agents with low confidence are more strongly influenced than agents who are already confident (cf. Matthes et al. 2010). This relationship is symmetrical: firm opinions are harder to erode, which can be argued on the basis of cognitive dissonance (Festinger 1957) and selective exposure (Knobloch-Westerwick 2014).

3.3 Experimental design

With the model fixed, it can now be connected to our research questions of how communities affect opinion expression and the formation of a spiral of silence.

In summary, the model has the following parameters:

1. the total number of agents,
2. a randomized initialization method of the *willingness to self-censor* Φ_i for the agents,
3. a randomized initialization method of the *confidence* $\hat{c}_i(t)$ for the agents at time t = 0,
4. the ratio of agents holding the positive opinion to agents holding the negative opinion,
5. the number of communities in the network,
6. the intra- and inter-community density of the communities, controlled by parameters m and ρ_{out} of the network model.

These parameters can be varied to measure how they affect the observable properties of the model over the course of a simulation. Since we aim to study opinion expression, and specifically the emergence of a spiral of silence, we first have to define what we consider a spiral of silence in our model. Again, we take inspiration from Ross et al. (2019), where the ratio of agents expressing their opinion to silenced agents was examined—unsilenced agents were also distinguished into those

belonging to the majority or minority opinion, based on all agents expressing their opinion. We say that a spiral of silence occurs in case "most" agents of one of the two opinions become silenced whereas the agents with the other opinion are almost all expressing their opinion.

The following (virtual) experiments consist of varying the model parameters while observing the dependent variables and thus interpreting the relationship between community structure and opinion expression. Note, however, that only the last two of the six model parameters are directly related to community structure while the others are indirect results of the modelling process in the context of the spiral of silence. The parameters 1–4 are therefore simply fixed to sensible values, while we make sure that their choice does not affect the results we obtain when varying parameters 5 and 6 (see the following section). We fix the number of agents to 1000 and initialize the willingness to self-censor and confidence by drawing from a uniform distribution in [0,1] independently for each agent. Each agent is independently assigned either the positive or the negative opinion (with a 50% probability of each). Since the opinion is also assigned independently of community membership, the distribution of positive and negative opinions is close to equal in each of the communities. While communities in real networks often exhibit homophily, and thus agents in the same communities should have more similar opinions, we chose not to model this explicitly as it would make it hard to identify which results are due to the network structure itself, or due to the fact that agents hold more similar opinions if they are in the same community.

Next to decide is how to vary the community-related parameters of the network model to study the research questions. Recall that, specifically, the goal is to examine if and how the fragmentation of a network into communities leads to an increased resilience against a spiral of silence. This can be investigated by varying the number and inter-community density of the communities generated by our network model. Note, however, that Ross et al. (2019) found that increasing the density of the networks, lead to a stronger spiral of silence effect. To account for this effect and study only the influence of the different community structures we have to make sure that the overall density of all generated networks stays constant.

Accordingly, for the first experiment we generate networks with 10 communities and vary the inter-community density ρ_{out} passed to our model. Then we study the relative size of the minority opinion among all agents expressing their opinions. Increasing the inter-community density without also changing the intra-community density, would make the overall network denser, leading to a higher synchronisation and a stronger spiral of silence. To deal with this problem we do not vary the inter-community density directly but change the m parameter of the Barabási-Albert model used to generate every community. The inter-community density is then chosen as a function of m such that the overall density stays constant.

In the second experiment we study the effect of the number and size of communities on the ability of a minority to keep expressing their opinion. To this end, we generate networks evenly partitioned into a varying number of communities (2–10). The overall density of the networks is kept constant by modifying, in this case, the inter-community density accordingly.

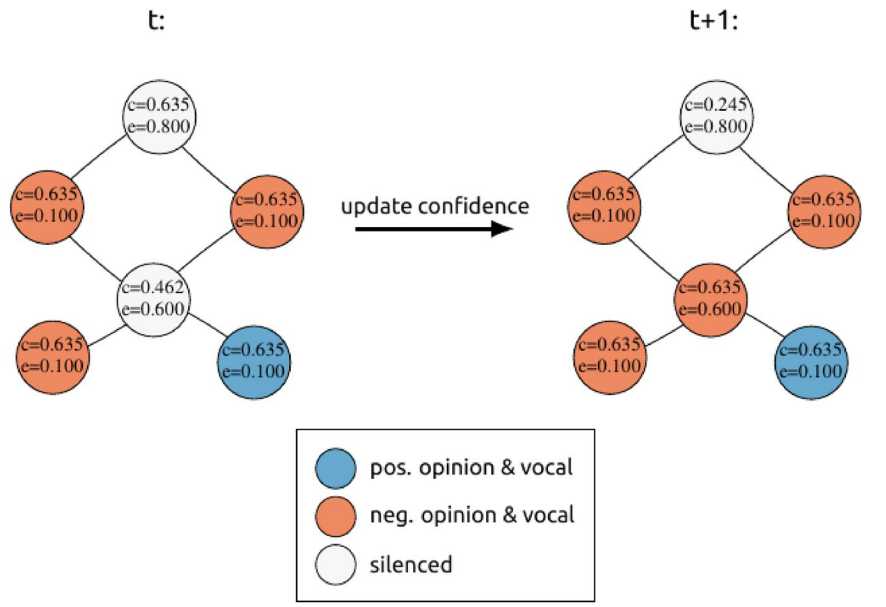

Fig. 2 Example confidence update visualisation generated using the program for running the simulations. The c and e values represent an agent's confidence and willingness to self-censor, respectively. The colors indicate the current state of an agent. In the example, we see the silenced agent in the middle gain confidence because in its neighborhood its opinion (−) is more prevalent than the (+) opinion. As a result of the increased confidence, the agent starts expressing its opinion again

3.4 Validation

An important step when working with agent-based models is validation. It is meant to guarantee that a model is an accurate representation of the studied real-world process. The following validation steps are based on the validation frameworks by Sargent (2013) and Klügl (2008) and include manually assessing visual animations of the model, studying degenerate edge cases, ensuring replicability of results in multiple runs, reproducing known results of Ross et al. (2019), and a sensitivity analysis of the input parameters.

Since our implementation allows for visual inspection during simulation runs, the validation process was started by comparing the agent interactions in very small model instances step by step to the expected behaviour, described in Sect. 3.2 (cf. Fig. 2). We also studied edge cases such as setting the willingness to self-censor to zero and making sure that no agents were ever silenced, or that in a model instance where all agents hold the same opinion, agents would over time all be expressing their opinion and not be silenced.

The experiments in the following section were always run multiple times to check that we had enough runs to get stable statistical results. Since we are relying on the agent behaviour described by Ross et al. (2019), we reproduced some of their results without bot agents by replacing our SBM network model with a simple Barabási–Albert model.

Another means of validation is running a sensitivity analysis, that is, examining if small changes in the model input parameters lead to vastly different outcomes. The underlying motivation is that the real-world model parameters can typically not be quantified perfectly, introducing a variability in the model inputs. This would make a very sensitive model less useful for predicting events in the real world. We used the one-factor-at-a-time (OFAT) method of sensitivity analysis described in (ten Broeke et al. 2016), varying the parameters described in the previous section one at a time while holding the other parameters constant, and validated that any variation in the outcomes was relatively small and that there were no critical points at which the behaviour changed extremely. Naturally, the first parameter, the number of total agents in the model, affected the absolute size of the factions (i.e., agents with positive and negative opinions, speaking or silenced agents). However, the relative sizes stayed more or less the same, except for very small instances of 50 agents or less. Varying the randomized initialization method for the agents' willingness to self-censor and the confidence (e.g., using uniform distributions in [0,2], [0,5] and [0,10], or exponential distributions with mean 1, 5 and 10) had almost no effect on the final stable state and thus the outcome of the experiments. This seems to be because a few steps into a simulation run, the confidence values adapt based on the values of their surrounding agents, a behaviour that was also observed by Ross et al. (2019).

The model was most sensitive with respect to changes in the distribution of agents' opinions. As described above, a 50:50 distribution was used in the experiments, where each agent was equally as likely to hold a positive or negative opinion. When we deviated from this equal distribution of opinions in simulations, we found that it became much harder for the minority opinion not to be silenced, even when there are only loosely connected communities. This is because we initialise agents' opinions independently across the network and so every community would also reflect a skewed global distribution making it likely that the more frequent opinion dominates in every community. However, while the size of a speaking minority shrinks when the opinion distribution deviates from 50:50, the trend displayed in Figs. 3 and 4 is still visible for distributions up to 30:70, after which the size of the speaking minority becomes essentially zero. We conclude that the results of our model apply in situations where the minority opinion is held by at least roughly 30% of people but caution should be exercised before generalising results to situations with smaller minorities.

The parameters 5 and 6, related to the community structure of the networks, are varied as part of answering the research questions and the results are described in the next section.

Finally, note that Ross et al. (2019) observed that the overall density of the network affects the strength of the observed spiral of silence. In a dense network, the high connectivity between agents seems to foster quick synchronization and no minority opinions are expressed anymore. As a result, and as already mentioned in the previous section, the overall network density was held constant when varying the community structure of the networks.

The external validity of a model is the ability that model results directly translate to scenarios observed in the real-world. In a best-case scenario, external validity can be tested by letting the model reproduce known empirical findings in the domain

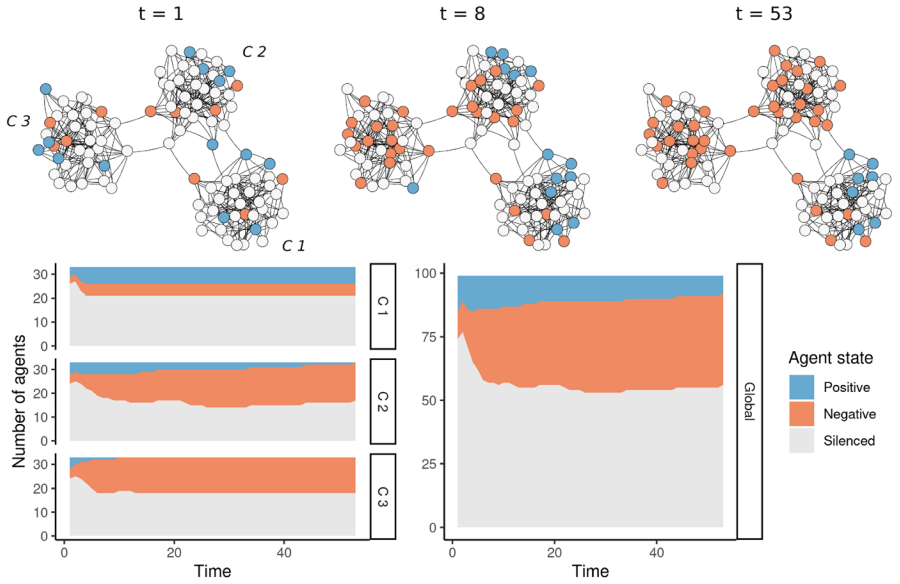

Fig. 3 Visualization of a simulation run of a single model instance. At the top, the model's state at three different points in time is drawn (start, after 8 steps, after the stable state is reached). The stacked area plots at the bottom display the distribution of agent states over time

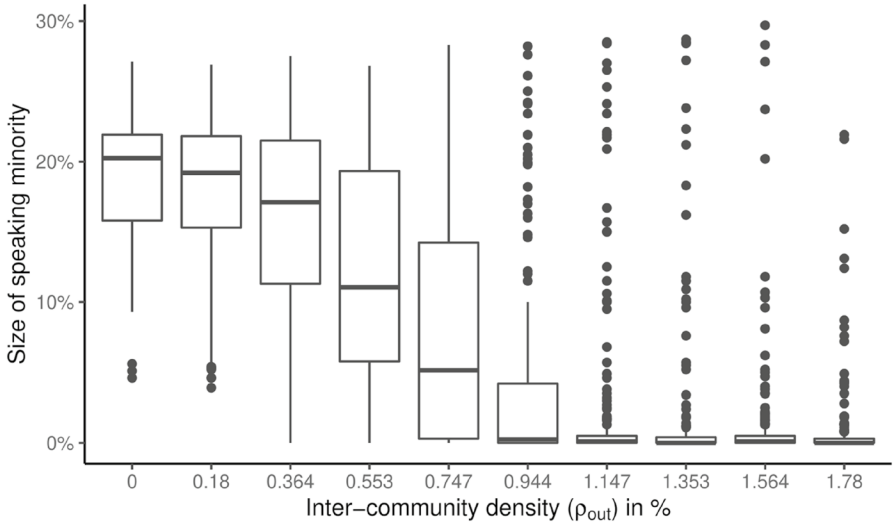

Fig. 4 The effect of inter-community density on the ability of the minority to keep expressing their opinion. The horizontal lines inside the boxes represent the median, while the upper and lower boundaries of the boxes are the 25th and 75th percentiles, respectively. The upper and lower whisker extend from the top or bottom of the box to the highest or lowest value, respectively, but no further than 1.5 times the box height. The points represent outliers

of interest, this is sometimes called predictive validity (Sargent 2013). To validate the present study, however, such empirical research would likely involve a large-scale survey asking participants for their opinions on a particular topic and asking if they are expressing that opinion publicly. Moreover, since we explicitly focus on the effect of community structure on the spiral of silence a comparable empirical study would also have to study multiple communities, and in the best case also quantify their mutual influences. There are various survey studies on the spiral of silence, see for example Glynn et al. (1997) for an overview. However, these are mainly concerned with verifying that the main mechanism of the spiral of silence actually exists, namely people self-censoring in face of a perceived opposing opinion climate. While there are some studies on the spiral of silence that explicitly mention communities (Salwen et al. 1994; Carter Olson and LaPoe 2017), they mostly focus on few, separate communities and not the interaction between multiple of them.

4 Results

Before we present the results of the experiments discussed in the previous section, we would like to give a better intuition on how the different modelling decisions work together. To this end, Fig. 3 visualises the simulation of a single model instance from time $t = 1$ to when the stable state is reached. The model consists of 99 agents uniformly distributed among 3 communities connected to each other only by a few connections. Initially, both opinions are expressed more or less equally in all communities. Over time, however, in communities "C 2" and "C 3" the negative opinion starts to dominate while any agents with positive opinions become silenced. In "C 1", a stable state, with some agents expressing positive and some expressing negative opinions, is reached. By the end, more than half of the agents are silenced. Because "C 2" and "C 3" are dominated by agents expressing negative opinions, the global distribution of expressed opinions is also heavily favoured towards negative opinions. The following experiments examine this behaviour for larger instances, with varying model properties, and averaged over a large number of runs.

Figure 4 shows the results for the first experiment, where the network contained $r = 10$ equally-sized communities, and the parameter m of the Barabási–Albert model was varied in $\{1, ..., 10\}$, while keeping the overall density of the networks on average constant at $\rho \approx 1.8\%$. This implied the corresponding variation of inter-community density displayed on the x-axis. The y-axis shows the percentage of agents that openly express the minority opinion, i.e., the opinion openly expressed by fewer agents compared to the other opinion. The displayed plot visualizes results of 500 randomized runs per configuration, 5000 runs overall.

We omit the percentage of agents expressing the majority opinion as well as silenced agents because the size of silenced agents stayed relatively constant and every increase in the majority opinion is reflected as a decrease in the minority opinion.

As expected, the speaking minority is strongest for $\rho_{out} = 0$, with 20% of the minority opinion still expressing their opinion. This is unsurprising because for $\rho_{out} = 0$ the communities are disconnected and the spiral of silence process develops

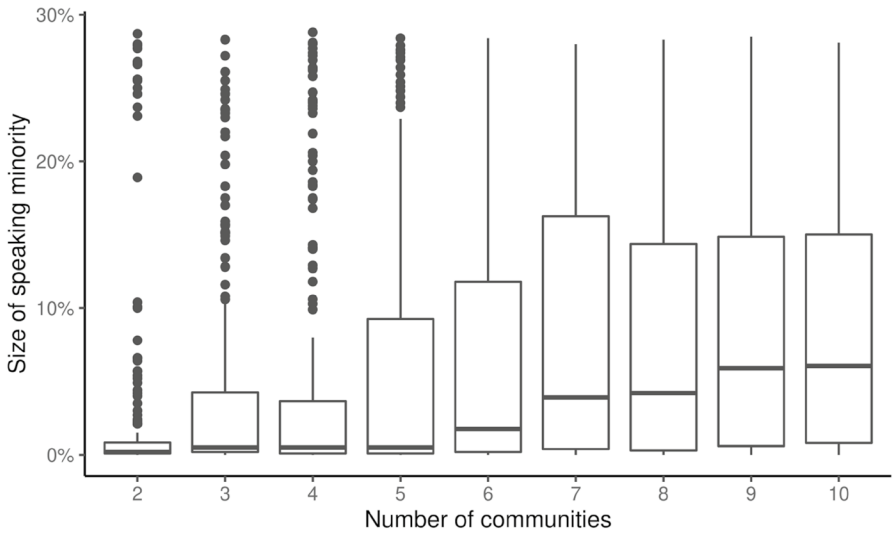

Fig. 5 The effect of the number of communities on the ability of a minority to keep expressing their opinion. The boxplot representation is the same as in Fig. 3

separately for each community. For 10 communities there is a high chance that there are communities in which the global minority opinion is dominating and not silenced. Since the communities are not connected to each other, such local majorities will not be silenced and are registered as part of the global minority expressing their opinion. This effect of completely disconnected communities seems to wear off when approaching an inter-community density of 1%. Further increasing the inter-community density, the ratio of the speaking minority stabilises at 1.5%, a value close to the one reported by Ross et al. (2019) for a network without communities at approximately $m = 6$. This to be expected, as the overall density in our networks matches a density in a network without communities generated only by the Barabási–Albert model for m between 5 and 6.

Figure 5 shows the result of the second experiment. Here, the number of equally-sized communities r was varied in $\{2, ..., 10\}$, while keeping the overall density of the networks on average constant. The parameter of the Barabási–Albert model was fixed at $m = 5$. The displayed boxplot visualizes results of 500 randomized runs per configuration, 5000 runs overall.

Intuitively a network with fewer, but larger communities should behave more similarly to a network without communities than a network fragmented into many smaller communities. Random disbalances of the agents' properties in a community can influence the dominating opinion in that community. In case of more communities, the probability that there will be communities with a differing minority/majority opinion compared to the overall network are higher.

From Fig. 5 it is apparent that a higher fragmentation into more, but smaller communities leads to more agents expressing a minority opinion. In the case of two communities, on average only 2.5% of agents belonging to the minority opinion

were still expressing their opinion at the end of the runs. When the networks consist of 10 equally-sized communities, up to 10% of agents holding the minority opinion are still openly expressing it. Together with the fact that the overall density of the network was kept constant, this seems to indicate that a fragmentation into more, smaller communities is beneficial for minorities to keep expressing their opinions and not be silenced by the majority.

5 Discussion

The present study examines how community structure affects the formation of public opinion, following the assumptions of the spiral of silence theory.

As a first result, we find that a high number of relatively small communities leads to a situation in which the minority opinion is still expressed by a larger part of the total population, compared to a scenario with a small number of large communities. In the former situation, entire subcommunities exist which have "local" majority opinions, undeterred by the fact that the global consensus is the opposite. Whether one views these small subcommunities in a positive light, as safe spaces in which minority opinions are still allowed to flourish, or negatively, as echo chambers of radicalisation, is open to interpretation. In the context of market segments, this result explains situations in which some markets lose interest in a product, as may happen in the airline industry if the flight shame movement continues to grow in the Western cultural sphere. Central nodes that are connected to many individuals have a greater influence on the opinions of others (van Eck et al. 2011). Furthermore, the influence of opinion leaders in a political context could be shown in the study of Twitter communities on the 2016 U.S. presidential election of Clinton and Trump, where certain opinion leaders led to a political homogeneity of the communication of communities (Guo et al. 2020). Thus it can be argued that our results of these small communities might be led by opinion leaders and their minority opinions. As Wu et al. (2015) reported, the frequency of connections to other nodes can lead to a convergence between communities, but this requires a uniform activation of all users and not only those users who exist as interfaces between the communities. Due to the fact that the connections to the individual communities are dependent on a few agents, they may not be in close contact with the opinion leaders and therefore are unlikely to be influenced (Liu 2007).

The second key result is that the more interconnected these communities are, the more likely a "global" spiral of silence is to emerge again. If the division of the network into communities creates "safe spaces" for minority opinions, a high degree of interconnectedness negates this effect. In other words, the more consumers from different markets communicate with each other, the more likely a spiral of silence is to emerge on a global scale. According to a related result by Sohn (2019), such a "global" spiral is also likely to occur in the case of mass media spreading a homogeneous opinion to a large part of the population. In an age of increasing global interconnectedness, in which information technology allows consumers to post their opinions on the internet for the entire world to see, this result would seem to predict an increasing homogenisation of consumer

opinion. However, “global” here refers to a spiral of silence encompassing the entire network of, in this case, 1000 actors. As Sohn (2019) points out, a truly world-wide spiral of silence is unlikely to occur, since the social network in neither simulation should be seen as an approximation of the social network of the 7.7 billion people in the world population, but rather the social network of some population of interest.

Several other studies reported results that are comparable to our findings. Wu et al. (2015) investigated different network topologies, one of which consisted of a network split into two communities. Similar to us, they found that “the number of silencers grows as the degree of coupling increases”. However, while we seem to replicate some of their results, they used a very different agent behaviour to simulate the spiral of silence process. In particular, they chose a single agent as the source of the initial opinion propagation then spreading to the rest of the network, and introduced an “immunity” that can keep agents from being silenced. The survival of minority opinions in the presence of sufficient modularity (i.e., community structure) is also a central result of Banisch and Olbrich (2019)’s model. Their approach shares with ours the distinction between opinion and opinion expression and it also relies on a positive/negative feedback mechanism not unlike those found in the spiral of silence theory, where agents are reinforced (or not) in their opinions by those around them. However, in Banisch and Olbrich’s model, agents are selected uniformly at random from the population and forced to express their opinions; silence is not an option. Since this is one of the defining features of the spiral of silence theory, Banisch and Olbrich’s results, while similar to ours, are the consequence of fundamentally different assumptions. In a direct comparison with both Wu et al. (2015) and Banisch and Olbrich (2019), the contribution of our research is to show that our model of the spiral of silence theory provides an alternative explanation for similar results.

When interpreting the results, the spiral of silence model needs to be distinguished from other models where the similarities are more superficial. The classical bounded confidence models such as the Deffuant model and the Hegselmann and Krause model show how opinions change over time in a continuous opinion value, but they do not show how confident agents feel about expressing their opinions and are therefore not convenient for modelling the processes of the spiral of silence. In the opinion dynamics model of Ye et al. (2019), which was inspired by the Friedkin-Johnsen model, a discussion process is simulated in which individuals adjust their private and expressed opinions in the network through the social influence of peer pressure. Here, variables such as stubbornness, resilience, individuals’ opinions are taken into account, which in a very dense network leads to quickly reach a "steady-state of persistent disagreement". However, these results are difficult to compare with our current study, since communities are not explicitly considered and the simulations focus on smaller numbers of agents, in contrast to our goal of simulating opinion dynamics in the large-scale online context. Although the results of Stern and Livan (2021) do not shed light on the spiral of silence, they do provide insights into opinion dynamics and show how opinions are distributed among communities in the network when they are created using the stochastic block model. The results of the study show that it is more difficult

for networks with closed communities to reach a common consensus when many different opinions exist, although the conceptualisation of what constitutes an opinion is rather different in their model and it lacks the distinction between opinion and opinion expression.

In terms of practical implications, companies can learn from the findings of this study. As described in Sect. 2, online opinion formation is a crucial factor for business success. Analysing the potential impact of community size, number and interconnectedness reveal several implications for strategic decision making within a company. For instance, establishing distinct communities for specific target markets, such as countries or products, could reduce the danger of fast-spreading negative opinions in case of an evolving corporate crisis. The management and interaction with customers can be used to establish partnerships with users of distinct communities, leading to more control on discussed topics on social media (Etter and Vestergaard 2015). In this context, the silence of a company on a discussed topic can have a negative impact on the opinion climate, and thus, on the business success (Stieglitz et al. 2019). Furthermore, the findings suggest that several smaller communities could act as a stabiliser for minority opinion expression. In the context of a corporate crisis, the minority expresses a positive opinion. Thus, the companies could maintain a positive opinion in those specific target markets. However, the establishment of distinct target markets, and therefore, communities, may not be sufficient enough in order to secure business success. Thus, the findings of this study implicate that companies should actively (1) observe and (2) manage, and (3) maintain the individual communities. Therefore, online community management may play a central role in a company's marketing planning. As a first step, potential communities need to be identified and continuously observed. Second, those communities should be actively managed, to this end, the company should communicate to customers and react to their feedback (e.g., customer co-creation). Third, the company should try to maintain a positive online opinion within the community by considering step two. To this end, the company might place corporate opinion leaders within the communities as communicators.

Of course, this study also faces distinct limitations. On the one hand, limitations of the spiral of silence theory have to be considered. Thus, the study models changes in the willingness to express one's opinion and not shifts in the held opinions themselves. On the other hand, the applied model is suitable for topics on which people have already formed their opinion and which do not change so quickly. Since the model gives each node in the network a 50% chance of being of the positive opinion, and a 50% chance of the negative opinions, the initial distribution of opinions within each community will rarely be exactly 50–50, but approach this in the long run. Such an approach is inappropriate to model a setting in which communities differ ideologically, such as an online community of car enthusiasts and an online community of environmentalists. However, in regard to realms such as general products or brand images, the applied model does allow concrete deductions for research and practice. Another limitation of our research is the empirical validation of the output of the model, considering that we do not have comparisons of theoretical foundations that deal with the spiral of silence theory linked to community structures. This problem of missing and non-existent data has already been addressed

in previous research (Fagiolo et al. 2007; Klügl 2008). For this reason, we took the approach of empirical data as input reference (Waldherr and Wettstein 2019), taking into account the empirical findings during the development of the model and their parameter settings. As Alvarez-Galvez (2016) indicates, using a connection of multiple techniques and data (real networks, media information, and survey methods), these agent-based models might be validated further in future research in order to gain a better understanding of the processes of opinion formation and their dynamics at different levels, beyond our validation efforts described in Sect. 3.4.

Analysing the findings provides foundations for several possible areas of future research. This could result in further insights about fields of application in which communities differ ideologically. Moreover, future research might distinguish between different types of actors within the network. Especially in the context of business success, actors such as opinion leaders and corporate influencers might play a special role. Therefore, the impact of opinion leaders, which may influence more or fewer people in relation to other actors, on global and community based spiralling effects could be examined.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abbe E (2017) Community detection and stochastic block models: recent developments. J Mach Learn Res 18(1):6446

Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47

Alvarez-Galvez J (2016) Network models of minority opinion spreading: using agent-based modeling to study possible scenarios of social contagion. Soc Sci Comput Rev 34(5):567–581

Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. Science 348(6239):1130–1132

Banisch S, Olbrich E (2019) Opinion polarization by learning from social feedback. J Math Sociol 43(2):76–103. https://doi.org/10.1080/0022250X.2018.1517761

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Barberá P, Jost JT, Nagler J, Tucker JA, Bonneau R (2015) Tweeting from left to right: is online political communication more than an echo chamber? Psychol Sci 26(10):1531–1542

Bikhchandani S, Hirshleifer D, Welch I (1992) A theory of fads, fashion, custom, and cultural-change as informational cascades. J Polit Econ 100(5):992–1026

Boutyline A, Willer R (2017) The social structure of political echo chambers: variation in ideological homophily in online networks. Polit Psychol 38(3):551–569. https://doi.org/10.1111/pops.12337

Bruch E, Atwell J (2015) Agent-based models in empirical social research. Sociol Methods Res 44(2):186–221. https://doi.org/10.1177/0049124113506405

Carter Olson CS, LaPoe V (2017) "Feminazis", "libtards", "snowflakes", and "racists": trolling and the Spiral of Silence effect in women, LGBTQIA communities, and disability populations before and after the 2016 election. J Public Interest Commun 1(2):116. https://doi.org/10.32473/jpic.v1.i2.p116
Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. Rev Mod Phys 81(2):591
Clifford P, Sudbury A (1973) A model for spatial conflict. Biometrika 60(3):581–588. https://doi.org/10.1093/biomet/60.3.581
Conover MD, Ratkiewicz J, Francisco M, Gonçalves B, Menczer F, Flammini A (2011) Political polarization on twitter. In: Fifth international AAAI conference on weblogs and social media
Czepkiewicz M, Árnadóttir Á, Heinonen J (2019) Flights dominate travel emissions of young urbanites. Sustainability 11(22):6340
Deffuant G, Neau D, Amblard F, Weisbuch G (2000) Mixing beliefs among interacting agents. Adv Complex Syst 3(01–04):87–98
DeGroot MH (1974) Reaching a consensus. J Am Stat Assoc 69(345):118–121
Epstein JM (2006) Generative social science: Studies in agent-based computational modelling, vol 13. Princeton University Press, Princeton
Erdős P, Rényi A (1959) On random graphs. I. Publ Math 6:290–297
Etter MA, Vestergaard A (2015) Facebook and the public framing of a corporate crisis. Corp Commun 20(2):163–177. https://doi.org/10.1108/CCIJ-10-2013-0082
Fagiolo G, Moneta A, Windrum P (2007) A critical guide to empirical validation of agent-based models in economics: methodologies, procedures, and open problems. Comput Econ 30(3):195–226. https://doi.org/10.1007/s10614-007-9104-4
Festinger L (1957) A theory of cognitive dissonance. Stanford University Press, Stanford
Flick C (2016) Informed consent and the Facebook emotional manipulation study. Res Ethics 12(1):14–28
Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174
Friedkin NE, Johnsen EC (1990) Social influence and opinions. J Math Sociol 15(3–4):193–206. https://doi.org/10.1080/0022250X.1990.9990069
Glynn CJ, Hayes AF, Shanahan J (1997) Perceived support for one's opinions and willingness to speak out: a meta-analysis of survey studies on the "spiral of silence." Public Opin Q 61(3):452–463
Grace-Farfaglia P, Dekkers A, Sundararajan B, Peters L, Park SH (2006) Multinational web uses and gratifications: measuring the social impact of online community participation across national boundaries. Electron Commer Res 6(1):75–101
Guerra P, Meira Jr W, Cardie C, Kleinberg R (2013) A measure of polarization on social media networks based on community boundaries. In: Proceedings of the international AAAI conference on web and social media, vol 7, no 1
Guo L, Rohde JA, Wu HD (2020) Who is responsible for Twitter's echo chamber problem? Evidence from 2016 U.S. election networks. Inf Commun Soc 23(2):234–251. https://doi.org/10.1080/1369118X.2018.1499793
Hayes AF, Glynn CJ, Shanahan J (2005) Willingness to self-censor: a construct and measurement tool for public opinion research. Int J Public Opin Res 17(3):298–323
Hayes AF, Uldall BR, Glynn CJ (2010) Validating the willingness to self-censor scale II: inhibition of opinion expression in a conversational setting. Commun Methods Meas 4(3):256–272
Hegselmann R, Krause U (2002) Opinion dynamics and bounded confidence models, analysis, and simulation. J Artif Soc Soc Simul 5(3)
Herring SC (ed) (1996) Computer-mediated communication: linguistic, social, and cross-cultural perspectives, vol 39. John Benjamins Publishing, Amsterdam
Holley RA, Liggett TM (1975) Ergodic theorems for weakly interacting infinite systems and the voter model. Ann Probab 3(4):643–663
Hopkins D (2016) Can environmental awareness explain declining preference for car-based mobility amongst generation Y? A qualitative examination of learn to drive behaviours. Transport Res Part A Policy Pract 94:149–163
IATA (2018) IATA Forecast Predicts 8.2 billion Air Travelers in 2037. https://www.iata.org/pressroom/pr/Pages/2018-10-24-02.aspx. Accessed 1 Dec 2019
Jiang W, Wu J (2017) Active opinion-formation in online social networks. In: IEEE INFOCOM 2017—IEEE conference on computer communications, Atlanta, pp 1–9. https://doi.org/10.1109/INFOCOM.2017.8057103
Kahn ME (2007) Do greens drive Hummers or hybrids? Environmental ideology as a determinant of consumer choice. J Environ Econ Manag 54(2):129–145

Kaiser C, Schlick S, Bodendorf F (2011) Warning system for online market research - Identifying critical situations in online opinion formation. Knowl Based Syst 24(6):824–836
Katz E, Lazarsfeld PF (1955) Personal influence; the part played by people in the flow of mass communications. Free Press, Glencoe
Klein D, Marx J, Fischbach K (2018) Agent-based modeling in social science, history, and philosophy. An introduction. Hist Soc Res 43(163):7–27
Klügl F (2008) A validation methodology for agent-based simulations. In: Proceedings of the 2008 ACM symposium on applied computing—SAC '08, p 39. https://doi.org/10.1145/1363686.1363696
Knobloch-Westerwick S (2014) Choice and preference in media use: advances in selective exposure theory and research. Routledge, New York
Liu FCS (2007) Constrained opinion leader influence in an electoral campaign season: revisiting the two-step flow theory with multi-agent simulation. Adv Complex Syst 10:233–250
Lorenz J (2006) Consensus strikes back in the Hegselmann–Krause model of continuous opinion dynamics under bounded confidence. J Artif Soc Soc Simul 9(1)
Ma S, Zhang H (2021) Opinion expression dynamics in social media chat groups: an integrated quasi-experimental and agent-based model approach. Complexity 2021:2304754. https://doi.org/10.1155/2021/2304754
Matthes J, Kimberly RM, Christian S (2010) A spiral of silence for some: attitude certainty and the expression of political minority opinions. Commun Res 37(6):774–800
McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27(1):415–444
Muchnik L, Aral S, Taylor SJ (2013) Social influence bias: a randomized experiment. Science 341(6146):647–651
Neubaum G, Krämer NC (2017) Monitoring the opinion of the crowd: psychological mechanisms underlying public opinion perceptions on social media. Media Psychol 20(3):502–531
Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
Noelle-Neumann E (1974) The spiral of silence: a theory of public opinion. J Commun 24(2):43–51
Oser J, Hooghe M, Marien S (2013) Is online participation distinct from offline participation? A latent class analysis of participation types and their stratification. Polit Res Q 66(1):91–101
Parker C (2019) Swedish climate activist Greta Thunberg is sailing to America amid a storm of online attacks. Washington Post. https://www.washingtonpost.com/world/2019/08/15/swedish-climate-activist-greta-thunberg-is-sailing-america-amid-storm-criticism/. Accessed 1 Dec 2019
Prior M (2007) Post-broadcast democracy: how media choice increases inequality in political involvement and polarizes elections. Cambridge University Press, Cambridge
Ridings C, Gefen D, Arinze B (2006) Psychological barriers: Lurker and poster motivation and behavior in online communities. Commun Assoc Inf Syst 18(1):16
Röchert D, Neubaum G, Ross B, Brachten F, Stieglitz S (2020) Opinion-based homogeneity on YouTube: combining sentiment and social network analysis. Comput Commun Res 2(1):81–108. https://doi.org/10.5117/CCR2020.1.004.ROCH
Ross B, Pilz L, Cabrera B, Brachten F, Neubaum G, Stieglitz S (2019) Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. Eur J Inf Syst (EJIS) 28(4):394–412
Rucinski T, Ringstrom A, Green M (2019) Airlines scramble to overcome polluter stigma as 'flight shame' movement grows. Reuters. https://www.reuters.com/article/us-airlines-iata-environment-analysis/airlines-scramble-to-overcome-polluter-stigma-as-flight-shame-movement-grows-idUSKCN1T4220. Accessed 1 Dec 2019
Salwen MB, Lin C, Matera FR (1994) Willingness to discuss "official English": a test of three communities. Journal Q 71(2):282–290. https://doi.org/10.1177/107769909407100203
Sargent RG (2013) Verification and validation of simulation models. J Simul 7(1):12–24. https://doi.org/10.1057/jos.2012.20
Sohn D (2019) Spiral of silence in the social media era: a simulation approach to the interplay between social networks and mass media. Commun Res. https://doi.org/10.1177/0093650219856510
Sohn D, Geidner N (2016) Collective dynamics of the spiral of silence: the role of ego-network size. Int J Public Opin Res 28(1):25–45
Stern S, Livan G (2021) The impact of noise and topology on opinion dynamics in social networks. R Soc Open Sci. https://doi.org/10.1098/rsos.201943
Stieglitz S, Mirbabaie M, Kroll T, Marx J (2019) Silence' as a Strategy during a corporate crisis—the case of Volkswagen's 'Dieselgate. Internet Res 29:4

Sunder S, Kim KH, Yorkston EA (2019) What drives herding behavior in online ratings? The role of rater experience, product portfolio, and diverging opinions. J Mark 83(6):93–112

Sznajd-Weron K, Sznajd J (2000) Opinion evolution in closed community. Int J Mod Phys C 11(06):1157–1165

Ten Broeke G, Van Voorn G, Ligtenberg A (2016) Which sensitivity analysis method should I use for my agent-based model? J Artif Soc Soc Simul 19(1):5

van Eck PS, Jager W, Leeflang PSH (2011) Opinion leaders' role in innovation diffusion: a simulation study: opinion leaders' role in innovation diffusion. J Prod Innov Manag 28(2):187–203. https://doi.org/10.1111/j.1540-5885.2011.00791.x

Velasquez A (2012) Social media and online political discussion: the effect of cues and informational cascades on participation in online political communities. New Media Soc 14(8):1286–1303

Waldherr A, Wettstein M (2019) Computational communication sciencel bridging the gaps: using agent-based modeling to reconcile data and theory in computational communication science. Int J Commun 13:24

Wasserman S, Faust K (1994) Social network analysis: methods and applications, vol 8. Cambridge University Press, Cambridge

Watts DJ, Dodds PS (2007) Influentials, networks, and public opinion formation. J Consum Res 34(4):441–458. https://doi.org/10.1086/518527

Weston R, Guia J, Mihalič T, Prats L, Blasco D, Ferrer-Roca N, Lawler M, Jarratt D (2019) Research for TRAN Committee—European tourism: recent developments and future challenges. European Parliament, Policy Department for Structural and Cohesion Policies, Brussels

Wilensky U, Rand W (2015) An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with Netlogo. MIT Press, Cambridge

Williams HT, McMurray JR, Kurz T, Lambert FH (2015) Network analysis reveals open forums and echo chambers in social media discussions of climate change. Glob Environ Chang 32:126–138

Wu Y, Du YJ, Li XY, Chen XL (2015) Exploring the spiral of silence in adjustable social networks. Int J Mod Phys C 26(11):1550125

Ye M, Qin Y, Govaert A, Anderson BDO, Cao M (2019) An influence network model to study discrepancies in expressed and private opinions. Automatica 107:371–381. https://doi.org/10.1016/j.automatica.2019.05.059

DOI: 10.17185/duepublico/76184
URN: urn:nbn:de:hbz:465-20220707-110419-7