UNIVERSITÄT
DUISBURG
ESSEN

Open-Minded

Instructional Awareness: A User-centred Approach for Risk Communication in Social Network Sites

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

Nicolás E. Díaz Ferreyra
aus
Santa Fe, Argentinien

1. Gutachter: Prof. Dr. Maritta Heisel
2. Gutachter: Prof. Dr. Nicole Krämer

Tag der mündlichen Prüfung: 19.02.2019

To my parents Susana and Carlos,
to their unconditional support,
and in return to all of those bedtime stories...

"In a world where everyone is overexposed,
the coolest thing you can do is maintain your mystery"

ANONYMOUS

Abstract

Often, users of Social Network Sites (SNSs) like Facebook or Twitter find hard to foresee the negative consequences of sharing private information on the Internet. Hence, many users suffer unwanted incidents such as identity theft, reputation damage, or harassment after their private information reaches an unintended audience. Many efforts have been made to develop preventative technologies (PTs) with the purpose of raising the levels of privacy awareness among the users of SNSs. Basically, these technologies generate interventions (i.e. warning messages) when users attempt to disclose private or sensitive information inside these platforms. However, users do not fully engage with PTs because they often perceive their interventions as too invasive or annoying. Basically, this happens because users have different privacy concerns and attitudes that should be considered when generating such interventions. In other words, some users are less concerned about their privacy than others and, consequently, are more willing to disclose private information without carrying much about the consequences. Therefore, PTs should incorporate *adaptivity* principles to their design in order to successfully nudge the users towards better privacy practices.

This thesis focuses in the development of an adaptive approach for generating privacy awareness in SNSs. Particularly, in the elaboration of software artefacts for communicating those privacy risks that may occur when disclosing private information in SNSs. Overall, this covers two main aspects: *knowledge extraction* and *knowledge application*. Artefacts for knowledge extraction include the data structures and methods necessary to represent and elicit risky self-disclosure scenarios in SNSs. In this work, privacy heuristics (PHs) are introduced as an alternative for representing such scenarios and as fundamental instruments for the generation of adaptive privacy awareness. Alongside, the artefacts corresponding to knowledge application comprise those methods and algorithms that leverage the information contained inside PHs to shape the corresponding interventions. This includes methods for estimating the risk impact of a self-disclosure act and mechanisms for regulating the content and frequency of warning messages. All of these artefacts collaborate with each other in a conceptual framework that this thesis calls Instructional Awareness.

Keywords: adaptive privacy, self-disclosure, awareness, social network sites, risk management, human-computer interaction

Kurzfassung

Für Nutzer von sozialen Netzwerken wie Facebook oder Twitter ist es meist eine Herausforderung die negativen Konsequenzen vorherzusehen, die das Teilen von privaten Informationen mit sich bringen. Daraus ergeben sich für die Nutzer häufig unerwünschte Vorfälle, wie beispielsweise Identitätsdiebstahl, Reputationsschäden oder Belästigungen – vor allem wenn deren private Informationen von einem nicht beabsichtigten Publikum konsumiert werden. In der Forschung wurden viele Bemühungen unternommen, um präventive Technologien zu entwickeln, sodass das Bewusstsein für die Privatsphäre der Nutzer gesteigert werden kann. Im Falle einer Veröffentlichung von privaten Informationen greifen diese Technologien verschiedene Interventionen auf (z.B. Warnmitteilungen). Dennoch sträuben sich Nutzer solch präventiven Technologien zu verwenden, da diese als zu invasiv und störend empfunden werden. Grund dafür sind die verschiedenen Datenschutzbedenken und Einstellungen zur Privatsphäre. In vielen Fällen ist die Sorge der Nutzer um ihre Privatsphäre gering, was zur Folge hat, dass sie häufig private Informationen veröffentlichen ohne sich über mögliche Konsequenzen Gedanken zu machen. Daher müssen bei der Entwicklung von Interventionen adaptive Mechanismen berücksichtigt werden, um auf die verschiedenen Datenschutzbedenken der Nutzer einzugehen. Dadurch können Nutzer in der Ausübung ihrer Privatsphäre unterstützt werden.

Diese Dissertation behandelt die Entwicklung eines adaptiven Ansatzes zur Förderung des Bewusstseins der Privatsphäre auf sozialen Netzwerken. Dabei wird vor allem auf die Elaboration von Softwareartefakten eingegangen, um die Risiken für die Nutzer in deren Privatsphäre zur kommunizieren. Daraus ergeben sich zwei Hauptaspekte: Wissensgewinnung und Wissensanwendung. Die Artefakte der Wissensgewinnung beinhalten Datenstrukturen und -Methoden, welche für die Erhebung und Repräsentation von riskanten Selbstenthüllungsszenarien notwendig sind. Diese Arbeit führt den Begriff der Privatsphäre-Heuristiken ein. Sie stellen eine Alternative zur Repräsentation der Selbstenthüllungsszenarien dar und können gleichzeitig als fundamentale Instrumente für die Generierung des adaptiven Bewusstseins genutzt werden. Gleichzeitig umfassen die Artefakte der Wissensanwendung diese Methoden und Algorithmen, welche die Informationen der Privatsphäre-Heuristiken enthalten, um die Interventionen zu entwickeln. Diese beinhalten Methoden zur Bewertung des Risikos bezüglich der Selbstenthüllung sowie Mechanismen zur Regulation des Inhalts und der Häufigkeit der Warnmitteilungen. Die verwendeten Artefakte kolla-

borieren in einem konzeptuellen Rahmen miteinander, welches in dieser Dissertation als „Instructional Awareness“ benannt wird.

Stichworte: adaptive Privatsphäre, Selbstenthüllung, Bewusstsein, soziale Netzwerke, Risikomanagemment, Mensch-Computer-Interaktion

Acknowledgments

This dissertation is the outcome of my work as a Ph.D. fellow at the University of Duisburg-Essen and as a member of the RTG "User-Centred Social Media". I want to thank on the first place my supervisor Prof. Maritta Heisel for her profound dedication and support during these years of work. Her knowledge, expertise, and guidance have been central not only for the content of this dissertation but also for my personal and professional development. Likewise, I want to acknowledge the help and support of my second supervisor Prof. Nicole Krämer, who nourished the interdisciplinary dimension of this work.

I would also like to acknowledge to all the members of the RTG who, through their vast expertise, helped me to improve different aspects of this thesis. Especially to Prof. Ulrich Hoppe for his constructive comments and suggestions that guided me in key areas of my research. Furthermore, I want to express my gratitude to my colleagues of the Software Engineering department who provided significant input to this work. Particularly, to my colleagues Dr. Azadeh Alebrahim, Angela Borchert, Ludger Goeke, Dr. Denis Hatebur, Rene Meis, Jens Leicht, Nelufar Ulfat-Bunyadi, Dr. Nazila Gol Mohammadi and Roman Wirtz for their constant support and assessment.

On the other hand, I want to extend my gratitude to those people who encouraged me to pursue a career in science and guided me in the previous stages to my Ph.D. In particular to Prof. Silvio Gonnet and Prof. Horacio Leone, who mentored me and guided me with great dedication and generosity at the very beginning of my academic career. Their supervision and advice at the Universidad Tecnologica Nacional in my hometown Santa Fe were essential for expanding my research interests and contributed significantly to my professional development. Likewise, I want to sincerely thank Prof. Ekkart Kindler from the Technical University of Denmark, who with extreme generosity offered me his time, expertise and assistance when I started pursuing a Ph.D. fellowship in Europe.

Last but not least, I want to thank my whole family for their unconditional support and help during this process. Furthermore, I want to extend my deepest gratitude to my friends and colleagues in Denmark, and Germany. This journey would not have been possible without their warm reception and hospitality during all these years away from my home country.

Included publications

Partial results of this dissertation have been published in:

- Nicolas E. Díaz Ferreyra and Johanna Schäwel. Self-disclosure in Social Media: An Opportunity for Self-Adaptive Systems. In *Joint Proceedings of the 22nd International Conference on Requirements Engineering: Foundation for Software Quality (REFSQ) Co-Located Events*, volume 1564 of *CEUR Workshop Proceedings*. REFSQ, CEUR-WS.org, March 2016.

- Nicolás E. Díaz Ferreyra, Johanna Schäwel, Maritta Heisel, and Christian Meske. Addressing Self-disclosure in Social Media: An Instructional Awareness Approach. In *Proceedings of the 2nd ACS/IEEE International Workshop on Online Social Networks Technologies (OSNT)*, pages 1–6. ACS/IEEE, December 2016.

- Nicolás E. Díaz Ferreyra, Rene Meis, and Maritta Heisel. Online Self-disclosure: From Users' Regrets to Instructional Awareness. In Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar Weippl, editors, *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 83–102. © Springer International Publishing, September 2017.

- Nicolás E. Díaz Ferreyra, Rene Meis, and Maritta Heisel. Towards an ILP Approach for Learning Privacy Heuristics From Users' Regrets. In Reda Alhajj, H. Ulrich Hoppe, Tobias Hecking, Piotr Bródka, and Przemyslaw Kazienko, editors, *Network Intelligence Meets User Centered Social Media Networks*, pages 187–197. © Springer International Publishing, August 2018.

- Nicolás E. Díaz Ferreyra, Rene Meis, and Maritta Heisel. Should User-generated Content be a Matter of Privacy Awareness? A position paper. In Kecheng Liu, Ana Carolina Salgado, Jorge Bernardino, and Joaquim Filipe, editors, *Proceedings of the 9th International Conference On Knowledge Management and Information Sharing (KMIS 2017)*, volume 3, pages 212–216. © INSTICC/SciTePress, November 2017.

- Nicolás E. Díaz Ferreyra, Rene Meis, and Maritta Heisel. At Your Own Risk: Shaping Privacy Heuristics for Online Self-disclosure. In *Proceedings of the 16th Annual Conference on Privacy, Security and Trust (PST)*, pages 1–10. IEEE, August 2018.

- Nicolás E. Díaz Ferreyra, Tobias Hecking, H. Ulrich Hoppe, and Maritta Heisel. Access-Control Prediction in Social Network Sites: Examining the Role of Homophily. In Steffen Staab, Olessia Koltsova, and Dmitry I. Ignatov, editors, *Social Informatics*, pages 61–74. © Springer International Publishing, 2018.

- Nicolás E. Díaz Ferreyra, Rene Meis, and Maritta Heisel. Learning from Online Regrets: From Deleted Posts to Risk Awareness in Social Network Sites. Submitted for publication, 2019.

Contents

Abstract v

Kurzfassung viii

Acknowledgments ix

Included Publications xi

Contents xiii

List of figures xv

List of tables xvii

Abbreviations xix

1 Research Outline 1

1.1 Introduction . . . 3

1.1.1 Online Self-disclosure . . . 3

1.1.2 Data Visceralization . . . 4

1.1.3 Preventative Technologies . . . 5

1.2 Research Motivation . . . 6

1.3 Research Questions . . . 7

1.4 Structure of the Dissertation . . . 10

2 Research Summary 13

2.1 Self-adaptation . . . 15

2.1.1 The MAPE-K Blueprint . . . 15

2.1.2 Constraint-based Modelling . . . 17

2.2 Instructional Awareness . . . 19

2.2.1 Architectural Model . . . 19

2.2.2 Privacy Heuristics . . . 20

2.2.3 Adaptation Loop . . . 21

2.3 Heuristics Elicitation . . . 22

2.3.1 Surveillance Attributes . . . 23

2.3.2 Regrets Identification . . . 24

2.3.3 Derivation Process 26
2.3.4 Audience Specification 28
2.4 Heuristics Application 31
2.4.1 Risk Estimation 31
2.4.2 Awareness Generation 32
2.5 Social Impact 34

3 Papers of the Dissertation **35**
3.1 Paper 1 37
3.2 Paper 2 39
3.3 Paper 3 41
3.4 Paper 4 43
3.5 Paper 5 45
3.6 Paper 6 47
3.7 Paper 7 49
3.8 Paper 8 51

4 Conclusion and Future Work **53**
4.1 Results 55
4.2 Discussion 59
4.3 Conclusion 61
4.4 Future work 62

Bibliography **65**

Appendix **74**

List of Figures

2.1 The MAPE-K Blueprint 16

2.2 Example of state constraint 18

2.3 Instructional Awareness System 20

2.4 IAS Adaptation Loop 22

2.5 Template for eliciting regrettable scenarios 25

2.6 Delete post interface 26

2.7 Active learning strategy for ACPMs 28

2.8 DT for predicting an unintended audience 30

2.9 ACL prediction trough CDAs 31

2.10 IAS's envisioned interface 33

4.1 Contributions map 59

List of Tables

2.1 Surveillance Attributes 23

2.2 Contingency Table 27

2.3 Disclosure Acceptance Matrix 29

3.1 Bibliographic information of Paper 1 37

3.2 Bibliographic information of Paper 2 39

3.3 Bibliographic information of Paper 3 41

3.4 Bibliographic information of Paper 4 43

3.5 Bibliographic information of Paper 5 45

3.6 Bibliographic information of Paper 6 47

3.7 Bibliographic information of Paper 7 49

3.8 Bibliographic information of Paper 8 51

4.1 Summary of outcomes 55

Abbreviations

ACL Access-control List.

ACPM Access-control Predictive Model.

CBM Constraint-based Modelling.

CDA Community-detection Algorithm.

CI Criticality Index.

CT Contingency Table.

DAM Disclosure Acceptance Matrix.

DPIA Data Protection Impact Assessment.

DT Decision Tree.

FTF face-to-face.

GDPR General Data Protection Regulation.

HWL Health Warning Label.

IAS Instructional Awareness System.

IASA Instructional Awareness Software Architecture.

ILP Inductive Logic Programming.

INL Information Nutrition Label.

ITS Intelligent Tutoring System.

KB Knowledge Base.

NFP need for popularity.

NLP Natural Language Processing.

NPO Nomprofit Organization.

PH Privacy Heuristic.

PHDB Privacy Heuristics Data Base.

PHeDer Privacy Heuristics Derivation Method.

PT Preventative Technology.

RQ Research Question.

SA Surveillance Attribute.

SDP Self-disclosure Pattern.

SNS Social Network Site.

UPDB User Performance Data Base.

1

Research Outline

The aim of this thesis is to generate awareness instruments for the users of Social Network Sites (SNSs) that can help them to foresee the negative consequences that may occur when revealing private information on the Internet. In this chapter the motivational aspects of this thesis are introduced together with a set of research questions that are addressed throughout this work. Particularly, the factors that contribute to self-disclosure in SNSs are analysed and discussed together with current approaches for privacy awareness. The absence of risk communication strategies in SNSs and the different privacy attitudes of the users are presented as the driving factors of this thesis.

1.1 Introduction

Privacy scholars have explored extensively the different factors that contribute to disclosing personal information in SNSs [7, 53, 41]. Likewise, efforts have been dedicated to understanding the way the users of these platforms put their privacy into practice [19, 59, 54]. In this section we discuss such factors together with the negative consequences of poor privacy practices on the Internet. Furthermore, we analyse state-of-the-art solutions that have been proposed for generating awareness in SNSs in order to identify areas of improvement.

1.1.1 Online Self-disclosure

The act of revealing personal information to others is commonly known as "self-disclosure". This mechanism, which is key for creating and maintaining social relations, take place in the real world and also on the Internet. Such is the case of SNSs like Facebook or Instagram, which allow their users to create a representation of themselves and, thereby, interact with each other [48]. That is, by adding, removing, editing and sharing media content (e.g. photos, text, videos, or links) on their profiles. Among the benefits of interacting in SNSs, the *social capital* is one of the most relevant ones [52, 20, 56, 5]. On a large scale, it refers to the advantages that we receive as a consequence of our social relationships. For instance, being friends of individuals with high social influences may contribute to our career development or give us access to special benefits. SNSs allow us to maintain multiple social connections simultaneously and, consequently, reinforce ones' social capital [20]. Other advantages of using SNSs include higher indices of psychological well-being such as self-steem and life satisfaction [52, 56, 5].

Like in the real world, links with friends and acquaintances in SNSs are reinforced as we disclose more personal information to them. However, the volume and type of content shared in SNSs is larger and more diverse than the one revealed offline [53, 2]. Up to some extent, this practice can be tracked to people's personality characteristics such as narcissism or need for popularity (NFP). Indeed, SNSs are ideal spaces for people with a high NFP since they allow to carefully plan ones' self-representation and, consequently, appear more social and popular [55]. However, self-representation is more challenging in SNSs than in the real world because the

size and composition of the audiences is hard to determine. That is, people are more aware of the information they share and the composition of the audience in a face-to-face (FTF) conversation than on the Internet [2]. Consequently, individuals use an "imagined audience" to guide their sharing behaviours which often does not correspond to the real one. This mismatch between the real and the imagined audience often derives in regrettable experiences for the users when their private information reaches an unintended audience.

1.1.2 Data Visceralization

Users of SNSs have reported in different studies and surveys to be concerned about their privacy [31]. However, the amount of personal information they reveal in these platforms contradict their claims. This mismatch phenomenon between peoples' privacy attitudes and their actual behaviour is known as the "privacy paradox"[3]. The privacy paradox, together with other self-disclosure models such as the *privacy calculus* [19], explain peoples' privacy behaviour through their attitudes and perceived benefits of sharing personal information. However, computers are social actors that can also influence our perception of information privacy and, consequently, our privacy behaviour. Basically, this is because private digital data is intangible and only perceived through the interfaces and physical materials of media technologies. Hence, such technologies modulate users' emotional perception and attachment towards their private information. Unfortunately, these technologies do not yet manage to take our perception to a *visceral* level. This is, making the tie between users' feelings and data visible, tangible and emotionally appreciable so they can perceive (in a visceral way) the impact of their disclosures.

Risk-awareness strategies are used often to inform people about the consequences of engaging with certain activities or consuming products or services. For instance, Health Warning Labels (HWLs) have become a standard for the communication of the risks of smoking and are required for the commercialization of cigarette packages in many countries [25]. However, when it comes to SNSs, users are not given much information about the privacy risks of online interaction such as cyber-bullying, identity theft, or reputation damage [14]. Moreover, when users give their consent for data collection and processing (i.e. when they accept the privacy policy), they receive very little (for not saying none) information about such risks. This lack of

information modulates the perceived severity of privacy risks in favour of information disclosure inside SNSs [47]. Furthermore, since SNSs profit themselves largely from user-generated content (e.g. by offering targeted advertisement services), low levels of risk awareness contribute (up to some extent) to the business model of these platforms [51].

1.1.3 Preventative Technologies

Privacy scholars have introduced several Preventative Technologies (PTs) that provide support (i.e. visual cues, warning messages or hints) to the users when they attempt to share private information in SNSs. Among the awareness strategies used by PTs, *nudging* is a well-known approach in which soft paternalistic interventions (i.e. information and guidance) are used to influence users' decisions towards safer and better choices [1]. In line with this approach, Wang et al. [57] developed three nudging strategies for Facebook consisting of (a) delaying the time before a post appears on the user's profile, (b) displaying visual cues about the post's audience, and (c) analysing and displaying the post's sentiment to the user. These nudges intervened when users attempted to post a message on Facebook allowing them to reconsider their disclosures and stay away from risky scenarios. Whereas this is an interesting approach for generating awareness, it can nevertheless result too invasive for the users when their particular privacy goals are not taken into account [14]. In other words, nudges can fail on engaging with those users who are less privacy-concerned or hold higher levels of privacy literacy when the frequency and intensity of the warnings is not regulated [49].

Another approach for privacy protection in SNSs are Access-control Lists (ACLs). Basically, these are data structures used to specify who can (or cannot) access certain pieces of private information in a particular context. Since creating and maintaining ACLs is often tedius, several strategies have been proposed for their automatic generation. Many of these Access-control Predictive Models (ACPMs) analyse *what* has been shared with *whom* in the past in order to recommend ACLs aligned with the users' privacy practices [14]. For instance, Misra and Such [35] developed an ACPM which considers previous sharing actions of the users on Facebook for the generation of ACLs. In this case, the generated ACL advises which type of content should be shared with whom depending on the type and strength of the online relationships

a user maintains with his/her Facebook friends [14]. Although machine learning approaches like this one show acceptable levels of accuracy, these methods still rely on the assumption that users have shared their content with the right audience in the past. Hence, the predicted ACLs not always fulfill the users' privacy preferences and goals particularlly for those users who show a great variation in their privacy behavior [36].

1.2 Research Motivation

People mostly regret having shared their private information in SNSs after risks are already materialized [58]. Moreover, they are likelly to protect their privacy more after experiencing an unwanted incident in person [8]. These regrettable experiences happen (in big part) because it is difficult for an average user to foresee the negative consequences of her privacy practices [58]. Awareness plays an important role in conducting the users towards a safer and preventive privacy behaviour. Particularly, nudging is a promising strategy for promoting best privacy practices among the users of SNSs. However, interventions should be generated according to the particular privacy goals of each user in order to achieve better levels of engagement. Moreover, they should incorporate risk information so users can decide what to share (or not) based on the negative consequences they may suffer.

Users can reduce their chances of living regrettable self-disclosure experiences by managing the audience of their publications. That is, building the corresponding ACL when the information they aim to publish can put their privacy into risk. For instance, a user who discloses a negative comment about her workplace should exclude her work colleagues from the audience in order to avoid negative consequences such as a wake-up call from a superior or even job loss. However, users are not highly motivated to create and maintain custom ACLs because, on top of its cognitive burden, it is hard for them to foresee the negative consequences of their disclosures. In other words, people are less encouraged to protect themselves when they are not aware of the risks they can suffer. Therefore, the adoption of ACLs should be motivated by risk awareness and supported by ACPMs that can satisfy the users' privacy goals and expectations.

1.3 Research Questions

As we have discussed, PTs should (i) put more emphasis on risk communication to motivate the users to protect themselves (ii) introduce ACPMs to relieve the users from the burden of building ACLs by hand, and (iii) intervene according to the privacy goals and expectations of the users to achieve higher engagement levels. Developing the software artefacts necessary to cover these features requires to address the following Research Questions (RQs):

RQ 1: Is there any architectural framework that could guide the development of adaptive PTs? As we discussed at the beginning of this chapter, adaptation is a key aspect for PTs. Such aspect is also required in a wide variety of systems that must operate under complex and changing environments and adapt their functionality accordingly [60]. Hence, many efforts have been made in the area of self-adaptive systems for providing methods and architectural blueprints to support and guide software engineers in their development [46]. One blueprint which is broadly adopted is the MAPE-K model introduced by IBM [26] which dissects self-adaptive solutions into five building blocks: *Monitor*, *Analyse*, *Plan*, *Execute* and *Knowledge*. This research question elaborates on the suitability of this architectural blueprint for achieving adaptivity in PTs. As it is shown in the upcoming sections of this dissertation, MAPE-K can be indeed leveraged to instantiate a software architecture for PTs oriented to nudge users in safer privacy practices. This motivates the following research question:

RQ 2: How should the architectural building blocks necessary to engineer adaptive PTs be instantiated? The prescribed MAPE-K architecture consists of generic building blocks that must be instantiated according to the requirements of each particular software project. In this case, it is expected that PTs provide personalized guidance and support to the users when disclosing private information inside SNSs. This functionality resembles in many aspects to the one of Intelligent Tutoring Systems (ITSs) which are used in learning environments [9]. Basically, an ITS is an expert system that recreates the intervention of human teachers providing personalized instructional content to students [37]. Up to a certain extent, users of SNSs can be considered as learners of privacy best practices. Therefore, the software modules of ITSs

can be adjusted and mapped to the MAPE-K blueprint in order to develop a specific architecture for adaptive PTs. The outcome of this research question is an architectural model for adaptive PTs called Instructional Awareness Software Architecture (IASA). Like in ITSs, one of the main components of IASA is a Knowledge Base (KB) used to track the user's performance and provide the right instructional feedback. This component motivates the next research question:

RQ 3: Which type of knowledge should be stored inside PTs and how can it be represented? The ultimate goal of PTs is to help users avoid regrettable self-disclosure scenarios. In consequence, PTs should be capable to identify these scenarios and the privacy risks associated with them. This suggests that these technologies should be endowed with a KB consisting of a collection of risky scenarios that can occur when revealing private information in SNSs. As we show in the following chapters, these scenarios can be modelled as patterns of information disclosure using principles of Constraint-based Modelling (CBM). Basically, CBM is a knowledge representation approach used in ITSs to describe the solution space of the tasks that students must solve [38]. This dissertation shows how CBM can be used for representing regrettable self-disclosure scenarios as Privacy Heuristics (PHs) that are used later on in the generation of risk awareness. Hence, the task of building a KB of PHs induces the following research question:

RQ 4: Which sources of information can be used to build a KB of regrettable scenarios and how can such information be retrieved? Privacy scholars have conducted several studies aiming to get better insight on the unwanted incidents that users experience when disclosing private information in SNSs [58, 7, 8]. Through FTF interviews and online questionnaires users have reported negative experiences such as cyber-bullying, identity theft and reputation damage. Therefore, a method could be elaborated to elicit PHs following the same approach. That is, a method that takes as input the experiences reported by the users for modelling the context in which such experiences take place. Moreover, as it is shown in a study by Wang et al. [58], deleted posts with private information can be considered as manifestations of regrets. Consequently, deleted posts can also be potential sources of PHs. In this thesis we introduce methods for the elicitation of PHs based on both, questionnaires and deleted posts.

As it is discussed in the next chapters, a PH models a regrettable scenario in terms of *risks*, *private information*, and an *unintended audience*. In other words, it represents the privacy risks that may occur when a piece of private information reaches a group of unintended recipients. Unlike the risks and the private information associated with a PH, the *audience* is a component that varies from individual to individual. This is because the network of connections of each user is composed by different people. Hence, the unintended audience of a PH must be personalized for each particular user. This leads to the following research question:

RQ 5: How can the audience of a PH be represented and personalized? Empirical studies have shown that online connections often represent the different social groups to which users belong in the real world [6, 32]. For instance, *family*, *workplace*, *nationality*, *religious beliefs* and *political affiliation* are often represented by the diversity of contacts that a user has inside a SNS. Alongside, these studies show that users often describe their audiences in terms of these social circles. Therefore, one could in principle describe the unintended audience of a PH in terms of abstract categories such as *family members* or *work colleagues* and refine them later on into personalized ACLs. That is, modelling the unintended audience of a PH as an ACL that represents a particular social circle. This strategy for audience representation is explored in the next chapters of this thesis. In order to overcome the issues of creating ACLs by hand we evaluate *active learning* strategies for their automatic generation. That is, ACPMs which generate ACLs with the participation of the users in a semi-automatic way.

As we mentioned, a self-disclosure scenario can result in one or more risks for the user. Therefore, a key aspect for generating awareness is the estimation of the risks that are associated with the PH that represents such scenario. Particularly, this motivates the following research question:

RQ 6: How can the risk of a self-disclosure scenario be estimated and used thereafter to communicate potential unwanted incidents? A risk is a characterization of the severity (i.e. a measure of the *consequence* and *frequency*) of an unwanted incident [33]. Consequently, one must have information about the likelihood of an unwanted incident and its impact. In principle, one could estimate the likelihood of an event by observing how frequently it occurs in a certain time frame. However, the consequence level of

an incident is subjective (i.e. varies from individual to individual), and the same incident can be perceived as *insignificant* by one user and *catastrophic* by others. Therefore, subjectivity is a factor that must be considered when estimating risks. In this thesis we propose an index that considers the subjective perception of the users for estimating the severity of risks. Alongside, we introduce an adaptive mechanism of risk awareness that uses such an index to regulate the amount of interventions (i.e. number of warning messages) to be generated by PTs.

As we mentioned, the ultimate goal of PTs is to protect the users from the risks that may occur when they disclose private information in SNSs. This goal is also shared and pursued by policy makers and Nonprofit Organizations (NPOs) who work on behalf of the users digital rights. Basically, this is done through the development of public policies which enforce SNSs to adopt good privacy practices. Hence, PTs could be used as vehicles for shaping public policies that promote the adoption of awareness mechanisms in SNSs. Therefore, the following research question arises:

RQ 7: How can PTs for risk communication and management support the public sector on expanding the users' privacy rights in SNSs? As we mentioned, SNSs do not offer much information about the risks of online interaction. Neither the privacy policy used to acquire the users' consent for processing their private information nor the platform layout include risk information. We believe that users should be empowered with risk awareness mechanisms while interacting in SNSs to make better privacy choices. Moreover, we believe that adaptive PTs for risk communication and management should be discussed and taken into consideration by policy makers to promote their adoption by social media platforms. Therefore, we discuss the value that this work has for the public sector and provide arguments towards more privacy-aware SNSs.

1.4 Structure of the Dissertation

The rest of this work is organized as follows. Chapter 2 introduces a summary of the publications included in this commutative dissertation and provides an overview of the most relevant aspects of this research. Particularly, section 2.2 elaborates on the

concept of self-adaptive software, and section 2.2 in the definition of an architecture for adaptive PTs. Likewise, section 2.3 introduces the software artefacts for the elicitation of regrettable scenarios, and section 2.4 those artefacts for the estimation of privacy risks together with the methods for the generation of personalized interventions. Next, chapter 3 presents the bibliographic information of the papers included in this thesis[1]. That is, information related to their publication outlet, type and publication status among others. It also shows the extent to which each paper addresses particular research questions together with the engineering methods applied in each case. Following, chapter 4 elaborates on the conclusions of this dissertation. In particular, section 4.1 maps the research questions to the different software artefacts introduced throughout this work, and section 4.2 discusses the strengths and limitations of our approach. Finally, we conclude in section 4.3, and recommend directions for future research in section 4.4.

[1]All the papers included in this cumulative dissertation can be found in the Appendix.

2

Research Summary

As we discussed in Chapter 1, nudging is a promising strategy for motivating the users of SNSs to reflect about their sharing behaviour. PTs aim to nudge the users in better privacy practices by intervening when they attempt to post content that may contain private information. Although this is a promising approach, we have identified issues related to the lack of personalization and risk communication that can hinder their adoption by the users. This dissertation introduces a set of software artefacts oriented to overcome these issues and, thereby, improve the effectiveness of PTs. In this chapter we provide an overview of the contributions of this dissertation and discuss how they answer the research questions introduced in Section 1.

2.1 Self-adaptation

Nowadays, personalization is an important aspect that must be considered when developing user-centred software solutions. From online shopping to online dating, personalized user experience is an important feature that systems should exhibit in order to engage more with their users. According to Riecken [45], "personalization is about developing customer loyalty by building meaningful one-to-one relationships; by understanding the needs of each individual and helping satisfy a goal that efficiently and knowledgeably addresses each individual's need in a given context" [50]. One of the main premises enclosed in this definition is that software solutions should adapt their functionality according to the goals of each user and some contextual information. In the particular case of PTs, this means that interventions should be generated according to the the privacy goals of each particular user in a certain disclosure scenario. Self-adaptation is a discipline that deals with the task of engineering systems that must achieve certain goals under changing conditions in the environment where they operate. One of the contributions of this thesis is an initial architecture for adaptive PTs based on a blueprint for self-adaptive software solutions.

2.1.1 The MAPE-K Blueprint

A distinctive characteristic of self-adaptive systems is a *control loop* mechanism that allows the system to adjust itself in response to internal changes (e.g. a failure) or changes in the environment (e.g. increasing network traffic). This mechanism consists of a set of processes necessary for monitoring the software system (the *self*) and its environment (the *context*) to detect significant changes, decide how to react, and act accordingly [46]. Therefore, self-adaptive software can be defined as a closed-loop system with feedback from the *self* and the *context*. This feedback loop is normally hidden or dispersed when the architecture of an adaptive system is documented or outlined. However, this feedback behaviour is a crucial feature that should be made explicit by the architectural blocks of a self-adaptive solution [4].

In order to guide the development of self-adaptive software, IBM engineers introduced an architectural blueprint that models control loop functionality explicitly. This blueprint dissects a control loop into four high-levels processes and a Knowl-

edge Base (KB) that interact with each other at runtime. As shown in Fig. 2.1, the control loop is implemented by an *autonomic manager* (i.e. the system to be) that operates in response to the information collected from a *managed resource* (i.e. the environment)[1]. The *Monitor* component provides the mechanisms to observe through *sensors* the different events or changes that take place at the managed resource. It is also responsible for processing, filtering, aggregating and reporting such information for future reference [4]. The *Analyze* component compares event data against patters in the KB to diagnose symptoms and thereby predict complex situations. Such symptoms are stored in the KB and used by the *Plan* component to determine the set of actions required to achieve a certain goal or objective. Finally, the *Execute* module performs the actions involved in a particular plan through its *effectors* to promote the necessary changes in the managed resource and, thereby, achieve the system's goals [29].

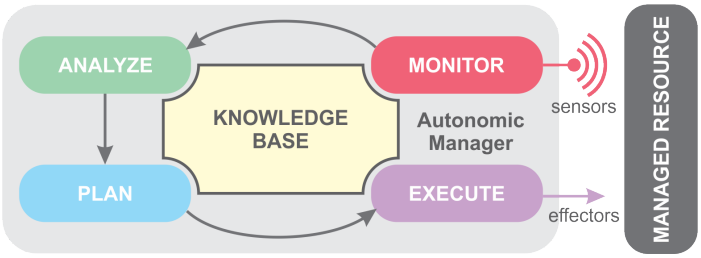


Figure 2.1: The MAPE-K Blueprint

This blueprint, commonly known as the MAPE-K (*Monitor*, *Analyse*, *Plan*, *Execute* and *Knowledge*), is used as an architectural reference model when a feedback loop is a distinctive characteristic of the system being built. However, while this model provides a good starting point for discussing feedback loop functionality, it does not provide any details on how to structure the raw data collected from the environment or how to represent the information stored inside the KB [4]. Moreover, it does not specify what data should be collected and what data should be stored. In other words, it gives a broad description of the main activities of a control loop without prescribing details for their implementation. Therefore, each particular software project must instantiate the MAPE-K according to its particular requirements and operational context [4].

As we described in Chapter 1, PTs are systems that operate in the context of SNSs

[1]Both, *autonomic manager* and *managed resource* are terms which are commonly used in the area of Autonomic Computing.

and aim to alert the users of these platforms about the risks that they may suffer when sharing private information on the Internet. If we analyse this scenario in terms of the components depicted in Fig. 2.1, we can map the PT to the Autonomic Manager and the user's account to the Managed Resource. That is, the PT can be considered as a self-adaptive system that monitors the user's account and intervenes when she attempts to publish a message with sensitive information. Therefore, the internal feedback loop of PTs could be defined in principle in terms of the generic components prescribed by the MAPE-K blueprint. For instance, through a *Monitor* module that collects information about the posts of the user, an *Analyse* module that compares such information against patterns of information disclosure stored in the KB, a *Plan* module that elaborates the corresponding warning message, and an *Execute* module that delivers the warning to the user[2]. Although this example is already a more concrete instance of the MAPE-K, it is necessary to conduct further refinements to define an implementable architecture of PTs.

2.1.2 Constraint-based Modelling

Some of the aspects that are not defined in the MAPE-K blueprint is how the different modules of the architecture cooperate with each other and how the information inside the KB should be represented. Although self-adaptation is considered a specific area of research nowadays, there is prior research in the areas of Control Theory and Artificial Intelligence that has addressed these issues to a large extent. For instance, knowledge representation and feedback generation are important aspects to be considered when developing Intelligent Tutoring Systems (ITSs). Basically, ITSs are artificial agents used in e-learning environments to provide personalized support and guidance to students. These systems are endowed with expert knowledge in a specific area (e.g. SQL [37]) that is used to mesure the progress of students and generate personalized instructional content [9]. Up to some extent, users of SNSs can be considered as learners of privacy best practices and PTs as an ITSs. Hence, state-of-the-art practices in ITSs can be adjusted and applied thereafter for engineering PTs.

Expert knowledge in ITSs is normally represented as a set of production rules that get activated at runtime. For instance, an ITS for SQL is likely to contain a KB with

[2]This suitability analysis of MAPE-K is introduced in detail in Paper 1.

if-then rules that describe the solution space of the queries that the students must solve. These rules are used to monitor the answers of the students and guide them to the correct solution. However, describing a solution space exhaustively can be a tough challenge and require a large number of production rules. Constraint-based Modelling (CBM) is a strategy for knowledge representation that aims to avoid over-specificity issues in ITSs. Basically, CBM suggests that all correct solutions to a problem satisfy a set of fundamental domain principles that should not be violated. For instance, the SELECT clause is mandatory in all SQL queries and, consequently, a constraint for all correct solutions. Therefore, domain knowledge can be expressed through a set of *state constraints* that model those basic domain principles that must be satisfied by any provided solution.

Problem State: The user writes a public post containing her new mobile number.

Constraint 1

Cr: *The user writes a comment containing personal contact information.*
Cs: *The post's audience must be anything but public.*

Figure 2.2: Example of state constraint

State constraints can be expressed as a pair of relevance and satisfaction tests *(Cr,Cs)* on a problem state, where each member of the pair can be regarded as a set of features or properties that a problem state must satisfy [39]. Therefore, domain knowledge in CBM is configured as a collection of state constraints of the form: "*if* relevance condition *Cr* is true, *then* satisfaction condition *Cs* had better also be true, otherwise something has gone wrong". For instance, let us assume that we are designing an ITS that, like in the case of PTs, aims to teach privacy best practices to the users of SNSs. A good privacy practice in SNSs is not to include private information such as address or phone number inside a public post. Hence, the KB of such ITSs should contain a constraint which represents this principle in order to evaluate the privacy behaviour of the user. For this particular constraint, *Cr* consists of checking if the user has disclosed personal information inside a post, and *Cs* verifies if the audience of such post is anything but public (as illustrated in Fig. 2.2). Therefore, when the user arrives to a problem state (i.e. a post) which

satisfies Cr (i.e. contains personal information) but not Cs (i.e. the audience is public), then the system intervenes with a warning message.

2.2 Instructional Awareness

So far, we have shown that MAPE-K offers a suitable framework for developing PTs, and CBM is a suitable approach for representing the knowledge that these technologies should incorporate. Therefore, it is possible to create an instance of MAPE-K for PTs in which knowledge is expressed as a collection of state constraints. An Instructional Awareness System (IAS) is a reference model for the design of PTs which combines self-adaptation and CBM. Similar to an ITS, IAS takes a pedagogical approach on privacy awareness in SNSs. That is, it considers the users of SNSs as learners of privacy best practices and aims for them to incorporate such practices to avoid regrettable experiences. For this, IAS intervenes with personalized warning messages when the users are about to publish posts with private information. These messages are generated when the user violates a *privacy best practice* which is known by IAS. Therefore, the KB of IAS consists of a collection of Privacy Heuristics (PHs) that are encoded as state constraints like in Fig. 2.2.

2.2.1 Architectural Model

The concept of IAS, and its corresponding Instructional Awareness Software Architecture (IASA), are contributions of this dissertation. A high-level view of IASA is shown in Fig. 2.3 (a detailed version can be found in Paper 2). As can be observed, the KB of IASA is divided into a Privacy Heuristics Data Base (PHDB) and a User Performance Data Base (UPDB). The first one corresponds to a collection of PHs that are evaluated when the user posts a message. Each of these PHs describe a pattern of private information that should not be disclosed to a particular audience. If the information disclosed inside a post violates any of these constraints, then IAS intervenes with a warning message. As we mentioned, interventions must be generated according to the privacy goals of each user. IAS regulates the frequency and intensity of the interventions based on a set of adaptation variables stored inside the UPDB. Such variables include how many times the user has ignored/accepted the system's warnings, and how often she discloses private information inside her

posts. Hence, the UPDB and the PHDB work closely together in detecting risky disclosures and guiding the users towards safer privacy practices.

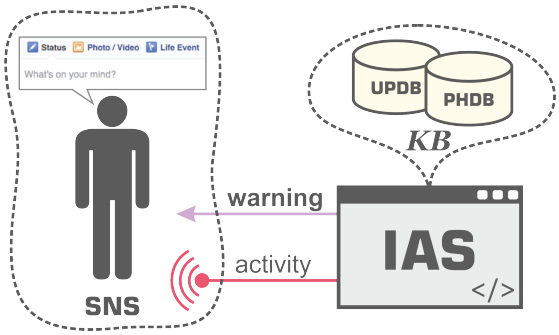


Figure 2.3: Instructional Awareness System

Although IASs resembles in many aspects the functionality of an ITSs, there are conceptual differences between them. The first one refers to the ultimate goal of IAS and the main purpose of an ITS. On one side, ITSs aim for all students to acquire the same amount of knowledge in a particular discipline (e.g. an SQL tutor aims for all the students to be capable of writing queries that are syntactically and semantically correct). As we mentioned, not all users have the same privacy goals or concerns. Therefore, the purpose of IAS is to provide the right instructional content so users can achieve a level of privacy knowledge which is aligned with their individual goals and concerns. The other distinctive aspect is the type of knowledge used by ITSs and IAS when shaping their interventions. On one hand, the domain principles that represent the state constraints of an ITS are pieces of *deterministic knowledge*. That its, if the student follows the domain principles that are represented by the state constraints, then she will eventually arrive to the solution of the problem. Conversely, the PHs that are part of IAS are pieces of *non-deterministic knowledge*. That is, if a user follows the domain principles represented by a set of PHs, then it is likely to reduce the chances of living a regrettable experience. Hence, a PH is a good privacy practice but following it does not guarantee the absence of future regrets or privacy harms.

2.2.2 Privacy Heuristics

PHs are key components of IASA and, consequently, of this thesis. The state constraint of Fig. 2.2 is a first approximation to a PH. Basically, it describes a regret-

table scenario in terms of two components: private information and an unintended audience. However, risk is also an important component of a PH since regrettable scenarios occur after the user has experienced an unwanted incident herself [58]. That is, after a risk has materialized. Hence, a PH can be resumed into a tuple <*PAs, Audience, Risk*> where *PAs* is a set of private attributes, *Audience* is a collection of recipients (e.g. Facebook *friends*), and *Risks* corresponds to the frequency and consequence of an unwanted incident. A PH models the relation between these three elements as *"The privacy Risk (i.e. consequence and frequency of an unwanted incident) associated with the disclosure of a set of PAs to an unintended Audience"*. For instance, let us consider a regrettable scenario in which a user is victim of *cyber-stalking* after revealing her phone number in a public post. Then, the corresponding PH for this scenario models the severity of cyber-stalking when a post containing one's phone number is revealed to a public audience in a SNS.

2.2.3 Adaptation Loop

Having the risks as a component of a PH allows IAS not only to communicate the user about the nature of the information she discloses (i.e. if it is private or not), but also the risks that may occur if such information reaches an unintended audience. As we explained in section 2.2.1, this can be done by querying the PHDB whenever the user aims to post a message. If there is a PH whose PAs correspond to the ones disclosed in the post, then IAS verifies that the post's audience is different to the described in such PH. If not, IAS generates a warning message using the risk information contained in the PH. For the phone number example, this message can be "Revealing your phone number to a public audience can result in episodes of cyber-stalking". This data flow, in which warning messages are generated using the information stored inside IAS's PHDB, corresponds to a process of *Knowledge Application*. That is, it assumes the existence of a collection of PHs inside IAS's PHDB. Hence, one must define the corresponding methods for *Knowledge Extraction* that allow IAS to incorporate PH to its PHDB. That is, to develop instruments for eliciting regrettable scenarios with the help of the users and shape thereafter the corresponding PHs.

Although knowledge extraction and application may seem to be independent of each other, these processes collaborate tightly in the generation of adaptive inter-

ventions. This becomes clearer when we analyse closer the role of the UPDB. On one side, PHs are used to shape the content of an intervention; however, the frequency of such interventions should be regulated to avoid irritating those users with low levels of privacy concerns. As we show in section 2.4, this can be done by observing the number of times a user ignores (or follows) IAS's interventions in a given period of time. Hence, the UPDB must collect this information after the knowledge of the PHDB is applied. As one can see, there is feedback between the application of PHs and the collection of the information necessary to regulate the frequency of IAS's interventions. Hence, the adaptation loop of IAS emerges from the synergy generated between the activities of knowledge application and extraction (as illustrated in Fig. 2.4).

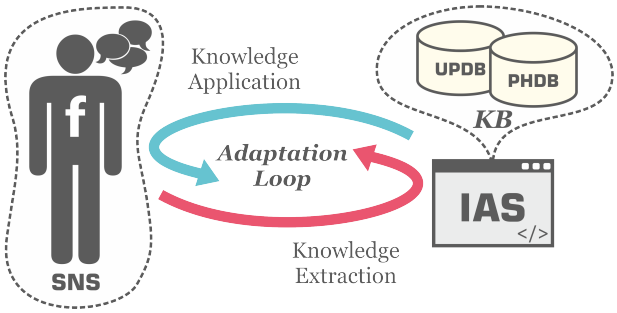


Figure 2.4: IAS Adaptation Loop

2.3 Heuristics Elicitation

As we mentioned, one of the main processes that take place in the IAS's adaptation loop is knowledge extraction. Basically, this covers those activities which are necessary for shaping PHs out of regrettable scenarios. In this section we provide an overview of the methods for the elicitation of PHs that are elaborated across the contributions of this thesis. All of these methods extract PHs out of regrettable experiences which are reported by the users. Hence, they consist of a number of steps which refine these reported experiences into the components of a PH (i.e. PAs, Audience and Risk). As we mentioned in chapter 1, modelling the audience of a PH is challenging since the network of connection of each user is composed by different people. Hence, we also describe the methods that we have elaborated for personalizing the audience component of PHs.

2.3.1 Surveillance Attributes

As we explained in section 2.2.1, to raise a warning, IAS checks if the information inside a post corresponds to the one represented in a PH. Posts in SNSs are pieces of unstructured data that are written in natural language. Therefore, it is necessary to abstract the PAs of a PH into a set of attributes that can be *automatically identified* through some Natural Language Processing (NLP) method. Hence, we must define a taxonomy of private attributes for representing the information contained inside regretable posts. There are several taxonomies of private attributes that can be found in the literature or in the body of regulatory frameworks for privacy protection. For instance, the General Data Protection Regulation (GDPR) [44] of the European Union specifies on its articles 4 and 9 which information must be considered as private or sensitive. It includes *name, identification number, location, and factors related to the physical, physiological, genetic, mental, economic, cultural or social identity* of a person. In principle, a taxonomy like this one could be adopted for our purpose. However, information that might not be personal per-se (e.g. a dynamic IP address) can be eventually linked to a person with the right contextual information [24]. Therefore, the content of a taxonomy depends largely on its purpose and the context of the information it aims to represent.

Dimension	**Surveillance Attributes**
Demographics	Age, Gender, Nationality, Racial origin, Ethnicity, Literacy level, Employment status, Income level, Family status
Sexual Profile	Sexual preference
Political Attitudes	Supported party, Political ideology
Religious Beliefs	Supported religion
Health Factors and Condition	Smoking, Alcohol drinking, Drug use, Chronic diseases, Disabilities, Other health factors
Location	Home location, Work location, Favourite places, Visited places
Administrative	Personal Identification Number
Contact	Email address, Phone number
Sentiment	Negative, Neutral, Positive

Table 2.1: Surveillance Attributes

In the case of IAS, a taxonomy of private attributes should cover those aspects of the users' personal information associated with-well known risks of online self-

disclosure. For instance, if a *job loss* incident is the consequence of revealing a negative comment about one's workplace, then the taxonomy should cover the attributes *work place* and *negative sentiment*. In a study conducted by Wang et al. [58] Facebook users described the information contained inside posts they later regretted. Hence, this study is a good starting point for building a taxonomy of attributes that are associated with risky self-disclosure scenarios. Such a taxonomy is illustrated in Table 2.1 and consists of a set of Surveillance Attributes (SAs) grouped around a number of high level categories called "self-disclosure dimensions". These categories, which keep a strong correlation with the regrettable scenarios reported in the study of Wang et al. [58], classify the SAs into *demographics, sexual profile, political attitudes, religious beliefs, health factors and condition, location, administrative, contact, and sentiment*. Developing the NLP tools for the automatic identification of these attributes goes beyond the scope of this thesis. Hence, we rely on the assumption that it is possible to describe posts and PHs as a set of SAs which can be identified by IAS.

2.3.2 Regrets Identification

The study of Wang et al. [58] not only helps us to understand under which circumstances people suffer from regrets, but also provides a strategy for eliciting regrettable scenarios. Basically, this study consisted of gathering a group of Facbook users and ask them "Have you posted something on Facebook and then regretted doing it? If so, what happened?". Users reported situations where posting about (a) alcohol and illegal drug use (b) sex (c) religion and politics (d) profanity and obscenity (e) personal and family issues (f) work and company and (g) content with strong sentiment, has resulted in regrettable experiences. As one can see, an empirical study like this one can be used as an instrument for eliciting regrettable scenarios and, consequently, PHs. This strategy, which is implemented by one of the methods for the elicitation of PHs introduced in this thesis, consists of gathering evidence on regrettable experiences through online questionnaires or face-to-face (FTF) interviews with the users. Such evidence, that must provide the information that is necessary for shaping PHs (i.e. PAs, Audience and Risk), can be captured using the template illustrated in Fig. 2.5. In this case, the template captures a scenario in which a user gets a wake-up call from her superior after posting a negative comment about her workplace. With this information one can start shaping the corresponding

PH that represents this particular scenario. This method, called Privacy Heuristics Derivation Method (PHeDer) is described in detail in Paper 3.

USER'S POST

"A typical day at the office. Lots of complaints and bad mood. Cannot wait for the day to be over...!"

Unintended Audience: The user's work colleagues, or superior.
Unwanted Incidents: Wake-up call from a superior
Perceived Consequence Level: HIGH.

Figure 2.5: Template for eliciting regrettable scenarios

Questionnaires and FTF interviews with the users are promising instruments for gathering regret evidence. However, a method for the elicitation of PHs which is based on these instruments can be expensive and inefficient due to the time that is necessary for recruiting a group of users, conducting a study, and process its results. Moreover, although the PHeDer method is useful for creating an initial collection of PHs, it must be applied periodically during the operational stage of IAS. This is necessary to capture new regrettable scenarios and, thereby, keep the PHDB updated. Hence, it is necessary to have an alternative method that allows the incorporation of PHs to IAS at any time without the burden of executing the PHeDer method periodically. For this purpose we introduced a method for the elicitation of regrettable experiences in SNSs based on deleted posts. As described by Wang et al. [58], users often delete their posts after living a regrettable experience. Therefore, deleted posts containing private or sensitive information are likely to be associated with a previous regret event. Under this premise, we developed a method that is triggered when a user deletes a post from her account. If the deleted post contains any of the SAs of Table 2.1, then it is likely to be associated with a regrettable event. Consequently, IAS requests the user who deleted the post to provide additional information through the interface illustrated in Fig 2.6. Like the template of Fig. 2.5, this interface elicits the unwanted incident suffered by the user, its perceived consequence level, and the unintended audience. If the post resulted in a regret, the user can specify the scenario herself using this interface and report it to IAS.

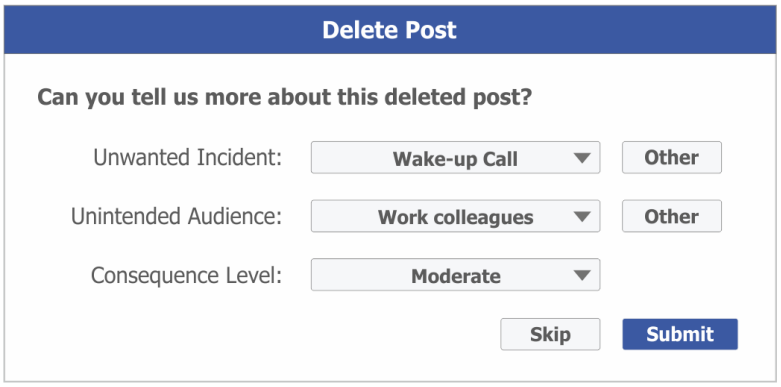


Figure 2.6: Delete post interface

2.3.3 Derivation Process

Overall, the derivation process of PHs can be summarized through the steps of Algorithm 1. Basically, this algorithm takes as input a regrettable scenario elicited using the PHeDer method (or through a deleted post) and refines it into a Self-disclosure Pattern (SDP)[3]. This pattern is a tuple $< SAs, Audience, UI >$ in which SAs is the set of SAs contained inside the post, $Audience$ is the untrusted circle of friends specified by the user, and UI corresponds to the reported unwanted incident. As we mentioned in section 2.3.1, it is possible to abstract the content of a post into a set of SAs. This is the first step into the derivation of a PH (line 1), and can be supported through NLPs methods such as regular expressions, named-entity recognition or deep learning algorithms[4]. Once the SAs inside the post are extracted, we proceed to build the SDP corresponding to the elicited scenario (line 2).

As one can observe, the elements of a SDP resemble the ones of a PH. However, whereas the third element of a PH is $Risk$, in an SDP it corresponds to the UI suffered by the user. This is because the estimation of a risk requires information related to the frequency of such UI, which is not available at this point. Hence, for a SDP to become a PH it is necessary to estimate the risk value of its UI. As mentioned in section 2.2.2, PHs (as also SDPs) are patterns that repeat themselves over time. Consequently, a SDP extracted from a user's report is likely to be extracted from the report of another user at some point. When this happens, regrettable experiences act as evidence of the same UI and, as shown in section 2.4.1, such ev-

[3]We have introduced SDPs to provide a consistent overview of this thesis' contributions. However, they are not referred by any of the papers included in this dissertation.

[4]For instance, Nguyen-Son et al. [42] applied support vector machines for the identification of private information in SNSs.

idence can be used later on in the estimation of privacy risks. Hence, after building the SDP corresponding to the user's report, we must check if such SDP is already contained inside the PHDB (line 3).

Algorithm 1 Heuristic elicitation pseudo-code

```
1: Extract SAs from the post
2: Build self-disclosure pattern (SDP)
3: if SDP does not exist in PHDB then
4:     Add SDP to PHDB
5:     Add new entry in the Contingency Table (CT)
6: else
7:     Update the corresponding entry in the CT
8: end if
```

If the SDP obtained from the user's report does not exist inside the PHDB, we must include it as a new entry (line 4). Likewise, we must add a new entry in a data structure called Contingency Table (CT) which organizes the information about the frequency of a SDP (line 5). Basically, a CT tracks the number of times a SDP has been reported with a particular impact value. For instance, according to the CT illustrated in Table 2.2, SDP1 has been reported 262 times: 10 as a scenario of catastrophic impact, 250 as major, and 2 as moderate. As one can observe, a CT aggregates the reports from all the users of IAS, and its information will help us to estimate the risk of a PH (as we show in section 2.4.2). Therefore, if the SDP is already contained inside the PHDB we must update its corresponding entry in the CT with the consequence value reported by the user (line 7). This process is fully described in Paper 8.

SDP	Catastrophic	Major	Moderate	Minor	Insignificant
SDP1	10	250	2	0	0
SDP2	30	150	20	0	0
SDP3	250	10	0	0	0
SDP4	0	50	100	5	0
SDP5	0	0	50	150	5

Table 2.2: Contingency Table

Both, PHs and SDPs, can be encoded as Horn clauses in Prolog. That is, one can implement a PHDB in which the relation between the elements of a PH (or a SDP) are modelled as logical predicates. This representation approach is used in Paper 3

and Paper 4. Particularly, when adopting this type of repressentation together with deleted posts as vehicles for the derivation or PHs, one can automate the first part of the process described in Algorithm 1. More preciselly, the step that correspond to the construction of the SDP from the user's report (line 2) can be automated using Inductive Logic Programming (ILP)[5]. This approach is described in detail in Paper 4.

2.3.4 Audience Specification

One of the aims of IAS is to nudge the users of SNSs into safer privacy practices. A good privacy protection strategy in SNSs consists of narrowing the audience of our posts. For the scenario of Fig. 2.5 this means that the user should exclude her work colleagues from the audience in order to minimize the chances of getting a wake up call from her superior. This practice can be supported by IAS through the recommendation of a personalized audience when users aim to disclose posts whose content is represented by a PH (or its corresponding SDP). Such audience recommendation consists of filtering out those contacts associated with the unintended audience of the corresponding PH. For this, the unintended audience of PHs must be specified as personalized ACLs. However, the elicitation methods we just introduced represent unintended audiences as generic social circles (i.e. as abstract categories such as as family members or work colleagues). Therefore, it is necessary to define a method to refine the unintended audience of PHs into personalized ACLs.

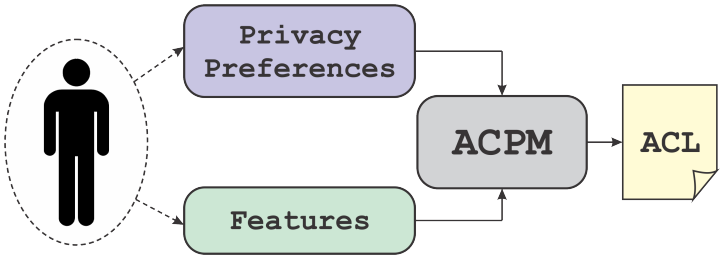


Figure 2.7: Active learning strategy for ACPMs

As we mentioned in section 1.1.3, ACPMs can be used to generate ACLs based on the users' previous privacy decisions (i.e. with whom they have shared private

[5]ILP is a subfield of machine learning in which logic programming is used to infer new knowledge (i.e. logical clauses) from a set of examples and some background knowledge [40].

information in the past). That is, they take the user's sharing behaviour as *predictor variables*, and particular instances of these predictors as *training examples* to predict ACLs. One of the drawbacks of this strategy is that it assumes that users have been careful when sharing their private information in the past. That is, it assumes that their posts have not resulted in a regrettable scenario and users have always shared their private information with the right audience. As we discussed in section 1.1.3, this can be a very strong assumption, especially for those users who show a large variation in their privacy behaviour. In order to reflect the true privacy preferences of the users, Fang and LeFevre [22] suggest an *active learning* approach in which users are interactively queried about their privacy preferences. In their approach, the *privacy preferences* of a user (i.e. the user's willingness or unwillingness to share information with each of her friends) are elicited by asking her to assign privacy labels to a representative sample of friends. These privacy preferences are then used together with a set of *features* (i.e. community structure and other information available inside the user's profile) to build an ACPM (as illustrated in Fig. 2.7).

Trusted Audience			
Friend	**Yes**	**No**	*example*
Bob		✗	e_1
Kate	✓		e_2
John		✗	e_3
Susan	✓		e_4
Bill		✗	e_5
Robin	✓		e_6
Marc	✓		e_7
Sally	✓		e_8

Table 2.3: Disclosure Acceptance Matrix

In order to follow an active learning strategy for the prediction of ACLs we introduce the concept of Disclosure Acceptance Matrix (DAM). Basically a DAM is a data structure used to elicit a sample of contacts that are considered members of an *untrusted* audience. As illustrated in Table 2.3, each row of a DAM consists of a tuple <*Friend, Allow*> where friend is someone from the user's contact list and allow is a *yes/no* value. With a DAM, a user can specify a sample of contacts to whom she is (not) willing to trust a particular piece of information (e.g. who should (not) access the comment of Fig. 2.5). Hence, the unintended audience of a PH can be elicited with a DAM (i.e. as an *untrusted* audience), and used afterwards for the

generation the corresponding ACL. This way, users get involved in the prediction process of the unintended audience and assumptions about their privacy preferences are avoided. One technique suitable for the prediction of ACLs are Decision Trees (DTs). A DT is a decision-support instrument which predicts the value of a target variable by learning simple decision rules inferred from a training set [30]. As we show in Paper 6, it is possible to train a DT that predicts the members of an unintended audience using the examples elicited with the DAM. As illustrated in Fig. 2.8, such a DT consists of a number of conditional tests over a set of *social attributes* (e.g. *workplace*, *gender*, and *location*). This hierarchy of rules inside the DT helps IAS to decide which Friend should be considered (or not) part of the unintended audience of a PH. Consequently, the DT can be used to filter the friend list of the user and, thereby, recommend a personalized ACL. This application of DTs for representing and predicting the unintended audience of PHs is detailed in Paper 6.

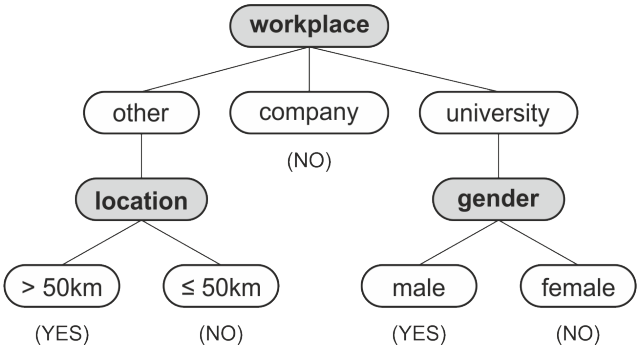


Figure 2.8: DT for predicting an unintended audience

Another strategy that is often used for the prediction of ACLs are Community-detection Algorithms (CDAs) which allow discovering clusters of densely connected nodes in a network. These algorithms can be applied to the user's *ego-network* (i.e. the network of connections between his/her friends) for finding clusters of well connected users [61, 36]. Thereby, an ACL can be created out of the community that best-fits the content of the DAMs. For instance, if a user Alice has selected her friends Bob, John and Bill as members of the unintended audience (i.e. like in the DAM of Table. 2.3), then this strategy searches for a community that groups together the most of these users. Let us assume that Alice's ego network can be clustered into three communities $C1$, $C2$ and $C3$, where Bob, John and Bill are grouped together in $C1$ (as illustrated in Fig. 2.9). Since community membership often indicates similarity between people, then one could expect that the rest of friends

inside community $C1$ are also ought to be considered as part of the unintended audience by Alice. Hence, an ACL consisting of all members of $C1$ is generated and recommended to Alice. This strategy is investigated and analysed in detail in Paper 7 through a simulation model of ego-networks.

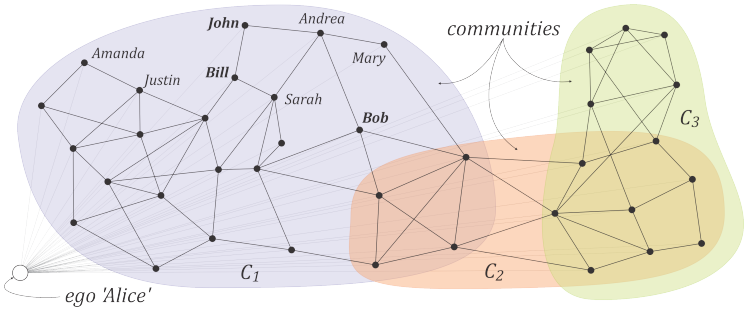


Figure 2.9: ACL prediction trough CDAs

2.4 Heuristics Application

So far, we have introduced the software artefacts that support the activities of knowledge extraction in IAS. As we mentioned in section 2.2.3, knowledge application covers those activities which use the information of the PHDB and the UPDB to shape IAS's interventions. These activities are triggered when IAS detects that the user attempts to disclose a post which contains a pattern of SAs that is represented by a PH. As mentioned in section 2.3.3, SDPs are the precursors of PHs and, therefore, can be used in the identification of potentially regrettable posts. However, in order to shape the corresponding intervention, IAS must evaluate the risk value of the unwanted incident represented by the SDP that the user violates with the content and audience of her post. Consequently, such SDP must be transformed into PH through the estimation of the risk value of its unwanted incident. Hence, risk estimation is a key step in the generation of warning messages. Following, we provide an overview of the software artefacts that we have engineered to support this process.

2.4.1 Risk Estimation

As mentioned in section 2.2.2, a risk is defined through the frequency and consequence of an unwanted incident. When eliciting a PH, the user provides information

about the consequence of the unwanted incident she suffered as result of a self-disclosure act. However, to estimate the frequency of such incident it is necessary to take into account more than just one observation of this particular event. That is, one cannot estimate that an incident will occur once a year or once a month only with a single piece of evidence. Moreover, the consequence level of an unwanted incident is subjective (i.e. varies from individual to individual) and the same incident can be perceived as *insignificant* by one user and *catastrophic* by others. Hence, one must have multiple evidence of an unwanted incident to estimate its frequency, and a metric that takes subjectivity into consideration to estimate its risk.

Previously, we have introduced PHs (and SDP) as patterns of regrettable scenarios that repeat themselves over the time. For instance, a user who posts *"I hate my job at this company but damn, it pays the rent! #keepcalm"* is revealing the same set of SAs as the ones being revealed in the post of Fig. 2.5. Moreover, like the user of Fig. 2.5, this other user can also suffer a wake-up call from her superior if the post is seen by her colleagues from work. As described in section 2.3.3, IAS keeps record of this phenomena in the CT with the purpose of estimating privacy risks. One way to do this is aggregating the information contained in the CT using a risk index. Basically, a risk index is a metric that combines instances of elementary risk evidence measured through quantitative or qualitative data [21]. In this thesis, we have adopted an index introduced by Faccinetti et al. [21] which aggregates the reported consequence values of an unwanted incident into a normalized risk value. Such value, called Criticality Index (CI), indicates a higher risk severity when it gets closer to 1 and lower as it approaches 0. The advantage of this index is that it can deal with consequence levels that are expressed in an ordinal scale such as *insignificant*, *minor*, *moderate*, *major* and *catastrophic*. A detailed description of CI together with its application to IAS is provided in Paper 8.

2.4.2 Awareness Generation

Algorithm 2 describes the main steps involved in the generation of adaptive warning messages. This process takes place whenever IAS detects that the user aims to post a message through her SNS account. In other words, IAS blocks the publication of the message in order to start the process described in Algorithm 2. The first step of this process consists of computing the set of SDPs that are *relevant* with regard

to the content of the post (line 1). That is, a set of SDPs whose SAs contain the ones included inside the post. For this, the content of the post is abstracted to a set of SAs and compared with the ones of each SDP inside the PHDB. Once this set of relevant SDPs is computed, IAS proceeds to determine which of these are *violated* by the targeted audience of the post (line 2). For this, the audience of the post is compared against the audience of each relevant SDP. If the audience of a relevant pattern contains the one of the post, then such a pattern is violated.

Algorithm 2 Awareness generation pseudo-code

```
1: Compute relevant SDPs
2: Compute violated SDPs
3: for each violated SDP do
4:     Compute risk
5: end for
6: Compute unacceptable risks
7: Communicate unacceptable risks
8: Update user's risk threshold
```

After computing the set of violated heuristics, IAS proceeds to estimate their respective risk values. Hence, IAS computes the CI corresponding to the unwanted incident of each violated SDP (lines 3 to 4). In other words, it transforms the violated SDPs into PHs. This is done using the information contained inside the CT introduced in section 2.3.3. As indicated in section 2.4.1, the CI generates a value between 0 and 1 which indicates how low or high is the severity of a risk. In order to define which unwanted incidents the user should be informed about, IAS needs to determine which risks are *unacceptable* for the user. For this, IAS compares the CI of each unwanted incident against the user's risk threshold (line 6). Such

Figure 2.10: IAS's envisioned interface

threshold is a value between 0 and 1 which synthesises the number of times the user has accepted (or rejected) IAS's suggestions in a given period of time. Those risks whose CI are above the user's threshold are unacceptable and, hence, should be communicated in the body of the warning message (line 7). Additionally, IAS can include an audience suggestion in the body of the warning message as shown in Fig. 2.10. After communicating the message to the user, IAS observes her behaviour (i.e. if she ignores the message or takes a preventive action) and updates the value of the threshold, accordingly (line 8). A detailed version of this algorithm is introduced in Paper 8[6].

2.5 Social Impact

The risks of revealing private information in SNSs can go way beyond job-loss and reputation damage. In fact, they can jeopardize the physical and mental integrity of those who suffer an unwanted incident [43]. This urges developers of media technologies and public policy makers to cooperate on behalf of the users' privacy rights. We believe that risk communication is a promising approach for supporting the users on making better privacy decisions when interacting in SNSs. Moreover, we believe that public policies are necessary to enforce the adoption of risk awareness mechanisms in social media platforms. In line with this, one aspect which is investigated in this thesis is the impact that IAS can have for the public sector. Particularly, if it is possible to develop public policies that promote the adoption of IAS in SNSs. For this, we have analysed a particular case in which law enforcement has been used to promote the adoption of awareness mechanisms: Health Warning Labels (HWLs) in cigarette packages. Despite the differences between tobacco consumption and the use of SNSs, HWLs demonstrate that law enforcement can be used to demand the inclusion of risk information in products or services [25]. Hence, HWLs are a success case which can provide plausible arguments towards the adoption of risk awareness systems like IAS in SNSs. This analysis is presented in Paper 5.

[6] A preliminary version of this algorithm is introduced in Paper 3. This first approach assumes that risk values are estimated with the help of a domain expert and not through a CI. Moreover, it assumes that the risk threshold of the user is given by her privacy attitude measured through privacy-related survey. These assumptions are dropped in the approach proposed in Paper 8.

3

Papers of the Dissertation

Due to the cumulative nature of this dissertation, the contributions of this work are distributed in a set of papers included in the Appendix. In this section we summarize the bibliographic information of each of these papers. This includes: authors and affiliations, abstract, publication outlet and publication type among others. Additionally, we indicate the engineering methods and techniques applied in each paper, and up to which extent each paper addresses the research questions introduced in Chapter 1. In general, a paper provides an answer to a particular research question. However, there are research questions that can be explored using different techniques and methods. Consequently, a research question which is addressed in one paper can be addressed in another one through a different technique. Furthermore, a paper can elaborate an aspect of a research question which contributes to its answer but requires further investigation. Hence, we represent with the symbol ● those research questions for which a paper offers an answer, and with the symbol ◓ those research questions it addresses but require further investigation. Each paper is self-contained in the sense that it has been published as a stand alone article. Hence, the contribution of one paper often appears as an assumption in another one. This is done in order to narrow the scope of each paper and systematically address all the research questions.

3.1 Paper 1

Title	**Self-disclosure in Social Media: An opportunity for Self-Adaptive Systems**
Author(s) & Affiliation(s)	Nicolás Emilio Díaz Ferreyra and Johanna Schäwel University of Duisburg-Essen, Germany RTG User-Centred Social Media https://www.ucsm.info/
Publication Outlet	2nd International Workshop on Requirements Engineering for Self-Adaptive and Cyber-Physical Systems (RESACS)
Publication Type	Workshop Paper
Publication Year	2016
Publication Status	Published
Keywords	social-media, self-disclosure, awareness, self-adaptive systems
Abstract	Users of Social Network Sites (SNSs) spend considerable amounts of hours per day exchanging (consuming or sharing) information and using services provided by such platforms. However, nothing comes for free. SNSs survive at the expense of the information that users' upload to their profiles, and the knowledge derived from their on-line behavior. Discovering hidden knowledge in social networks is a centerpiece in many personalized on-line services and ad-targeting techniques, and helps to make a SNS profitable. However, users seem not to be aware of this common practice and keep sharing content compulsively. Nevertheless, self-disclosure and over-exposition can have severe consequences and can put users' integrity into risk. In order to develop better information control and awareness systems, we believe that it is important to take into account the users' on-line habits and behavior. In this work we introduce an initial assessment of the different factors that contribute to self-disclosure in Social Media, and discuss the elements that a self-adaptive solution should consider to address this issue.
Addressed RQs	RQ1: ● RQ2: ◓
Applied Method(s)	Software Architectures for Self-adaptive Systems

Table 3.1: Bibliographic information of Paper 1

Reprinted with permission from Nicolas E. Díaz Ferreyra and Johanna Schäwel. Self-disclosure in Social Media: An Opportunity for Self-Adaptive Systems. In *Joint Proceedings of the 22nd International Conference on Requirements Engineering: Foundation for Software Quality (REFSQ) Co-Located Events*, volume 1564 of *CEUR Workshop Proceedings*. REFSQ, CEUR-WS.org, March 2016.

3.2 Paper 2

Title	**Addressing Self-disclosure in Social Media: An Instructional Awareness Approach**
Author(s) & Affiliation(s)	Nicolás Emilio Díaz Ferreyra, Johanna Schäwel, Maritta Heisel and Christian Meske University of Duisburg-Essen, Germany RTG User-Centred Social Media https://www.ucsm.info/
Publication Outlet	2nd The Second International Workshop on Online Social Networks Technologies (OSNT)
Publication Type	Workshop Paper
Publication Year	2016
Publication Status	Published
Keywords	social network sites, self-disclosure, constraint-based modelling, self-adaptive systems, intelligent tutoring systems
Abstract	Nowadays the information flowing across the different Social Network Sites (SNSs) like Facebook is highly diverse and rich in its content. It is precisely the diversity of the users' contributions to SNSs that makes these platforms attractive and interesting to engage with. Nevertheless, there is a high amount of private and sensitive information being disclosed permanently by these users in order to take full advantage of the services offered by such sites. Current privacy-protection approaches (like the one provided by Facebook) allow users to restrict the audience of their contributions and hide particular pieces of information; however, they are still far from being widely adopted and put proactively into practice. For this reason, we propose to analyze and address different aspects of online self-disclosure in Social Media from a pedagogical and self-adaptive perspective. In this work we introduce the architecture of an Instructional Awareness System (IAS) based on the MAPE-K blueprint for autonomic systems, and provide a definition of its feedback mechanism using principles of Constraint-Based Modeling (CBM).
Addressed RQs	RQ2: ● RQ3: ◓
Applied Method(s)	Software Architectures for Self-adaptive Systems, Constraint-based Modelling

Table 3.2: Bibliographic information of Paper 2

3.3 Paper 3

Title	Online Self-disclosure: From Users' Regrets to Instructional Awareness
Author(s) & Affiliation(s)	Nicolás Emilio Díaz Ferreyra, Rene Meis and Maritta Heisel University of Duisburg-Essen, Germany RTG User-Centred Social Media https://www.ucsm.info/
Publication Outlet	International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE)
Publication Type	Conference Paper
Publication Year	2017
Publication Status	Published
Keywords	social network sites, adaptive privacy, awareness, heuristics, risk analysis
Abstract	Unlike the offline world, the online world is devoid of well-evolved norms of interaction which guide socialization and self-disclosure. Therefore, it is difficult for members of online communities like Social Network Sites (SNSs) to control the scope of their actions and predict others' reactions to them. Consequently users might not always anticipate the consequences of their online activities and often engage in actions they later regret. Regrettable and negative self-disclosure experiences can be considered as rich sources of privacy heuristics and a valuable input for the development of privacy awareness mechanisms. In this work, we introduce a Privacy Heuristics Derivation Method (PHeDer) to encode regrettable self-disclosure experiences into privacy best practices. Since information about the impact and the frequency of unwanted incidents (such as job loss, identity theft or bad image) can be used to raise users' awareness, this method (and its conceptual model) puts special focus on the risks of online self-disclosure. At the end of this work, we provide assessment on how the outcome of the method can be used in the context of an adaptive awareness system for generating tailored feedback and support.
Addressed RQs	RQ3: ● RQ4: ● RQ5: ◓
Applied Method(s)	Constraint-based Modelling, Logic Programming, Risk Analysis

Table 3.3: Bibliographic information of Paper 3

3.4 Paper 4

Title	**Towards an ILP Approach for Learning Privacy Heuristics From Users' Regrets**
Author(s) & Affiliation(s)	Nicolás Emilio Díaz Ferreyra, Rene Meis and Maritta Heisel University of Duisburg-Essen, Germany RTG User-Centred Social Media https://www.ucsm.info/
Publication Outlet	4th European Network Intelligence Conference (ENIC)
Publication Type	Conference Paper
Publication Year	2017
Publication Status	Published
Keywords	adaptive privacy, self-disclosure, awareness, social network sites, inductive logic programming
Abstract	Disclosing private information in Social Network Sites (SNSs) often derives in unwanted incidents for the users (such as bad image, identity theft or unjustified discrimination), along with a feeling of regret and repentance. Regrettable online self-disclosure experiences can be seen as sources of privacy heuristics (best practices) that can help shaping better privacy awareness mechanisms. Considering deleted posts as an explicit manifestation of users' regrets, we propose an Inductive Logic Programming (ILP) approach for learning privacy heuristics. In this paper we introduce the motivating scenario and the theoretical foundations of this approach, and we provide an initial assessment towards its implementation.
Addressed RQs	RQ4: ●
Applied Method(s)	Inductive Logic Programming, Risk Analysis

Table 3.4: Bibliographic information of Paper 4

3.5 Paper 5

Title	**Should User-generated Content be a Matter of Privacy Awareness? A position Paper**
Author(s) & Affiliation(s)	Nicolás Emilio Díaz Ferreyra, Rene Meis and Maritta Heisel University of Duisburg-Essen, Germany RTG User-Centred Social Media https://www.ucsm.info/
Publication Outlet	9th International Conference on Knowledge Management and Information Sharing (KMIS)
Publication Type	Position Paper
Publication Year	2017
Publication Status	Published
Keywords	adaptive privacy, self-disclosure, awareness, social network sites, data visceralization
Abstract	Social Network Sites (SNSs) like Facebook or Twitter have radically redefined the mechanisms for social interaction. One of the main aspects of these platforms are their information sharing features which allow user-generated content to reach wide and diverse audiences within a few seconds. Whereas the spectrum of shared content is large and varied, it can nevertheless include private and sensitive information. Such content of sensitive nature can derive in unwanted incidents for the users (such as reputation damage, job loss, or harassment) when reaching unintended audiences. In this paper, we analyse and discuss the privacy risks of information disclosure in SNSs from a user-centred perspective. We argue that this is a problem of lack of awareness which is grounded in an emotional detachment between the users and their digital data. In line with this, we will discuss preventative technologies for raising awareness and approaches for building a stronger connection between the users and their private information. Likewise, we encourage the inclusion of awareness mechanisms for providing better insights on the privacy policies of SNSs.
Addressed RQs	RQ7: ●

Table 3.5: Bibliographic information of Paper 5

Reprinted by permission from INSTICC/SciTePress: Nicolás E. Díaz Ferreyra, Rene Meis, and Maritta Heisel. Should User-generated Content be a Matter of Privacy Awareness? A position paper. In Kecheng Liu, Ana Carolina Salgado, Jorge Bernardino, and Joaquim Filipe, editors, *Proceedings of the 9th International Conference On Knowledge Management and Information Sharing (KMIS 2017)*, volume 3, pages 212–216.

3.6 Paper 6

Title	At Your Own Risk: Shaping Privacy Heuristics for Online Self-disclosure
Author(s) & Affiliation(s)	Nicolás Emilio Díaz Ferreyra, Rene Meis and Maritta Heisel University of Duisburg-Essen, Germany RTG User-Centred Social Media https://www.ucsm.info/
Publication Outlet	16th International Conference on Privacy, Security and Trust (PST)
Publication Type	Conference Paper
Publication Year	2018
Publication Status	Published
Keywords	social network sites, adaptive privacy, awareness, heuristics, risk analysis, human-computer interaction
Abstract	Revealing private and sensitive information on Social Network Sites (SNSs) like Facebook is a common practice which sometimes results in unwanted incidents for the users. One approach for helping users to avoid regrettable scenarios is through awareness mechanisms which inform a priori about the potential privacy risks of a self-disclosure act. Privacy heuristics are instruments which describe recurrent regrettable scenarios and can support the generation of privacy awareness. One important component of a heuristic is the group of people who should not access specific private information under a certain privacy risk. However, specifying an exhaustive list of unwanted recipients for a given regrettable scenario can be a tedious task which necessarily demands the user's intervention. In this paper, we introduce an approach based on decision trees to instantiate the audience component of privacy heuristics with minor intervention from the users. We introduce Disclosure-Acceptance Trees, a data structure representative of the audience component of a heuristic and describe a method for their generation out of user-centred privacy preferences.
Addressed RQs	RQ5: ●
Applied Method(s)	Decision Trees, Risk Analysis

Table 3.6: Bibliographic information of Paper 6

3.7 Paper 7

Title	**Access-control Prediction in Social Network Sites: Examining the Role of Homophily**
Author(s) & Affiliation(s)	Nicolás Emilio Díaz Ferreyra, Tobias Hecing, H. Ulrich Hoppe and Maritta Heisel University of Duisburg-Essen, Germany RTG User-Centred Social Media https://www.ucsm.info/
Publication Outlet	10th International Conference on Social Informatics (SOCINFO)
Publication Type	Conference Paper
Publication Year	2018
Publication Status	Published
Keywords	homophily, preferential attachment, adaptive privacy, access-control prediction, social network sites
Abstract	Often, users of Social Network Sites (SNSs) like Facebook or Twitter have issues when controlling the access to their content. Access-control predictive models are used to recommend access-control configurations which are aligned with the users' individual privacy preferences. One basic strategy for the prediction of access-control configurations is to generate access-control lists out of the emerging communities inside the user's ego-network. That is, in a *community-based* fashion. Homophily, which is the tendency of individuals to bond with others who hold similar characteristics, can influence the network structure of SNSs and bias the users' privacy preferences. Consequently, it can also impact the quality of the configurations generated by access-control predictive models that follow a community-based approach. In this work, we use a simulation model to evaluate the effect of homophily when predicting access-control lists in SNSs. We generate networks with different levels of homophily and analyse thereby its impact on access-control recommendations.
Addressed RQs	RQ5: ●
Applied Method(s)	Social Network Analysis, Simulation

Table 3.7: Bibliographic information of Paper 7

3.8 Paper 8

Title	**Learning from Online Regrets: From Deleted Posts to Risk Awareness in Social Network Sites**
Author(s) & Affiliation(s)	Nicolás Emilio Díaz Ferreyra, Rene Meis and Maritta Heisel University of Duisburg-Essen, Germany RTG User-Centred Social Media https://www.ucsm.info/
Publication Outlet	27th ACM Conference on User Modelling, Adaptation and Personalization (UMAP)
Publication Type	Conference Paper
Publication Year	2019
Publication Status	Under Revision
Keywords	adaptive privacy, self-disclosure, awareness, social network sites, risk management
Abstract	Social Network Sites (SNSs) like Facebook or Instagram are spaces where people expose their lives to wide and diverse audiences. This practice can lead to unwanted incidents such as reputation damage, job loss or harassment when pieces of private information reach unintended recipients. As a consequence, users very often regret to have posted private information in SNSs and proceed to delete such content after these risks are materialized. Privacy scholars have developed different preventative technologies that raise awareness on the privacy risks of online self-disclosure. However, many of these approaches assume that information about risks is retrieved and measured by an expert in the field and, consequently, pass over this important aspect. In this work we introduce an approach that employs deleted posts as risk information vehicles to measure the frequency and consequence of unwanted incidents inside SNSs. In this method, consequence is reported by the users through an ordinal scale and used later on to compute a risk criticality index. We thereupon show how this index can serve in the generation of adaptive risk warnings.
Addressed RQs	RQ6: ●
Applied Method(s)	Risk Estimation

Table 3.8: Bibliographic information of Paper 8

4

Conclusion and Future Work

This section discusses the outcomes of this dissertation. For this, we analyse how the research questions introduced in chapter 1 are addressed in each paper of the dissertation. That is, up to which extent a paper addresses a particular research question and which are the main outcomes of each individual contribution. Furthermore, we relate these outcomes to relevant aspects of IAS such as its operational environment, adaptation phases and architecture. In addition, we analyse the advantages and points of improvements of our approach together with directions for future research.

4.1 Results

As we discussed in chapter 1, privacy as a self-disclosure issue is a multifaced and complex problem that cannot be addressed extensively from a single discipline. Moreover, it is an issue that has different nuances, each representing a wide variety of research challenges. This thesis has focused on the role that risk awareness plays when users reveal private aspects of their lives in SNSs. Particularly, in the role that media technologies have in the communication of those privacy risks that are related to regrettable self-disclosure scenarios that may take place in SNSs. For this purpose, we have we have elaborated on a set of Research Questions (RQs) in section 1.3 that have guided the development of the approach introduced in this dissertation.

Paper	RQ1	RQ2	RQ3	RQ4	RQ5	RQ6	RQ7	Main outcome
Paper 1	●	◓	○	○	○	○	○	MAPE-K for engineering adaptive PTs
Paper 2	○	●	◓	○	○	○	○	Conceptualization of IAS, IASA
Paper 3	○	○	●	●	◓	○	○	PHs, PHeDer method
Paper 4	○	○	○	●	○	○	○	PHs elicitation using deleted posts
Paper 5	○	○	○	○	○	○	●	Social impact of IAS
Paper 6	○	○	○	○	●	○	○	DTs for representing the audience of PHs
Paper 7	○	○	○	○	●	○	○	Prediction of ACLs using CDAs
Paper 8	○	○	○	○	○	●	○	Risk estimation of PHs

Table 4.1: Summary of outcomes

As shown in chapter 3, each RQ has been addressed in one or more papers, and each paper contributes to the development of one or more software artefacts (e.g. IASA, PHs). Table 4.1 provides a summary of the contributions of each paper in terms of RQs and outcomes. The contribution of a paper to a RQ is represented with a ● symbol when such paper *offers an answer* to the RQ, a ◓ symbol when the paper *elaborates an aspect* of a RQ that requires further investigation, and a ○ symbol when the paper *does not contribute* to a RQ. Thus, the RQs of this dissertation have been addressed through the different outcomes of each paper as follows:

RQ 1: Is there any architectural framework that could guide the development of adaptive Preventative Technologies (PTs)? As we discussed in section 2.1.1, the MAPE-K blueprint can be adopted as an architectural framework for engineering those adaptive features that PTs must exhibit in order to fulfil the individual privacy goals of the users. The adoption of this framework has helped us to reflect on the critical areas that must be addressed when developing user-centred solutions for risk awareness in SNSs. Since the suitability of MAPE-K was analysed in Paper 1, the outcome of this paper offers an answer to this particular RQ1. Moreover, the outcome of this paper already provides a high-level description of a MAPE-K instance for PTs (i.e. the role of each module in the generation of warning messages). Hence, it also elaborates an aspect of the next research question:

RQ 2: How should the architectural building blocks necessary to engineer adaptive PTs be instantiated? In section 2.2 we have defined IAS together with its corresponding architectural model IASA. Basically, IASA is an instantiation of the MAPE-K blueprint which is introduced in Paper 2. Hence, IASA and its corresponding Paper 2 offer an answer to this particular RQ2. Furthermore, the same paper introduces Constraint-based Modelling (CBM) as a suitable strategy for representing the information inside the Knowledge Base (KB) of IAS. That is, it proposes to encode self-disclosure scenarios as *state constraints* which are used thereafter by IAS for the generation of personalized warning messages (as described in section 2.1.2). Hence, this paper also elaborates an important aspect of the following research question:

RQ 3: Which type of knowledge should be stored inside PTs and how can it be represented? The introduction of CBM as a strategy of knowledge representation in IAS is a first attempt in answering this research question. Basically, a self-disclosure scenario modelled as a state constraint consists of a pair of relevance and satisfaction conditions *(Cr, Cs)* that are tested against the content of a post. However, the risk associated to a particular scenario is not modelled by any of these conditions. Hence, PHs were introduced as an alternative to state constraints and to model explicitly the risks of a self-disclosure scenario. As shown in Table 4.1, this approach is introduced as an answer to RQ3 in Paper 3. Furthermore, this paper introduces PHeDer, a method for extracting PHs out of regrettable scenarios reported by the users. Hence, Paper 3 provides also an answer to the following research question:

RQ 4: Which sources of information can be used to build a KB of regrettable scenarios and how can such information be retrieved? The PHeDer method introduces a strategy for shaping PHs out of regrettable scenarios that are reported by the users. This method uses FTF interviews and online questionnaires as instruments for collecting information about regrettable self-disclosure scenarios in SNS. As we described in section 2.3.2, this approach can be expensive and inefficient due to the time and effort required for conducting a study with a group of users. For this, we have introduced in Paper 4 an alternative method that also provides an answer to this RQ4. This method consists of using deleted posts as vehicles for reporting regrettable scenarios in SNSs. That is, users who delete posts with personal information are likely to have experienced a regrettable scenario themselves. Under this premise, the method asks the information necessary for shaping a PH to those users who delete a post containing private information.

The PHeDer method introduced in Paper 3 elaborates on the representation of the different components of a PH (i.e. SAs, Audience, and Risk). In the particular case of the Audience, it proposes a representation approach consisting of generic social circles (e.g. "family members" or "work colleagues"). Thus, Paper 3 elaborates an aspect of the following research question:

RQ 5: How can the audience of a PH be represented and personalized? As we discussed in section 2.3.4, the fundamental limitation of using generic social circles is that the composition of a circle varies from individual to individual. For instance, the network of contacts representing a user's "family members" is unlikely to be the same as the one from another user. Hence, social circles must be refined into personalized Access-control Lists (ACLs) as we indicate in section 2.3.4. Two papers of this dissertation provide an answer to this RQ5. The first one is Paper 6 which introduces Decision Trees (DTs) for predicting the composition of an ACL out of a set of examples provided by the user. The second is Paper 7 which explores Community-detection Algorithms (CDAs) for creating personalized ACLs out of the emerging communities inside the users' ego-networks. This last paper explores this strategy using a simulation model of scale-free networks.

One of the main aspects of risk awareness consists of estimating the risk value of the unwanted incident associated to a PH. However, none of the papers we have mentioned so far address specifically the following research question:

RQ 6: How can the risk of a self-disclosure scenario be estimated and used thereafter to communicate potential unwanted incidents? As mentioned in section 2.4.1, estimating the risk value of an unwanted incident is often a challenging task. Basically, this is because the impact of an unwanted incident can be perceived differently by different groups of users. For instance, some users may perceive "reputation damage" as a *catastrophic* event whereas others may perceive it as a *minor* issue. Hence, we propose to estimate risks through a Criticality Index (CI) that takes this subjectivity aspect into account. Such an index (introduced in section 2.4.1) aggregates information reported by the users regarding the perceived consequence level of an unwanted incident. The value generated by this index is a central component of the mechanism discussed in section 2.4.2 for the generation of personalized warning messages. This mechanism together with the corresponding CI are elaborated in Paper 8. Hence, this paper offers an answer to this RQ6.

An important aspect of this thesis is the implications that this work has to the users' privacy rights. That is, how does it contribute to the public sector in shaping public policies that enforce social media platforms to put more emphasis on privacy awareness. Consequently, in Paper 5 we have addressed the following research question:

RQ 7: How can PTs for risk communication and management support the public sector on expanding the users' privacy rights in SNSs? Throughout this work, we have discussed the importance that risk awareness has for our privacy behaviour inside SNSs. We have also highlighted in section 2.5 the absence of risk information in both, the body of privacy policies and on the layout of social media platforms. Under this scenario, we believe that the approach introduced in this work can contribute to engineer more privacy-aware social environments on the Internet. Particularly, through the incorporation of IAS as a risk awareness feature in SNSs. This can be promoted through a public policy for SNSs that considers users as *consumers* that have the right to be informed about potential threats to their privacy. This proposal is introduced in Paper 5 as an answer to this RQ7. Moreover, it has been discussed in the lightning session "Awareness by design: On the roads to self-determination" at the 2018 Internet Governance Forum of the United Nations[1].

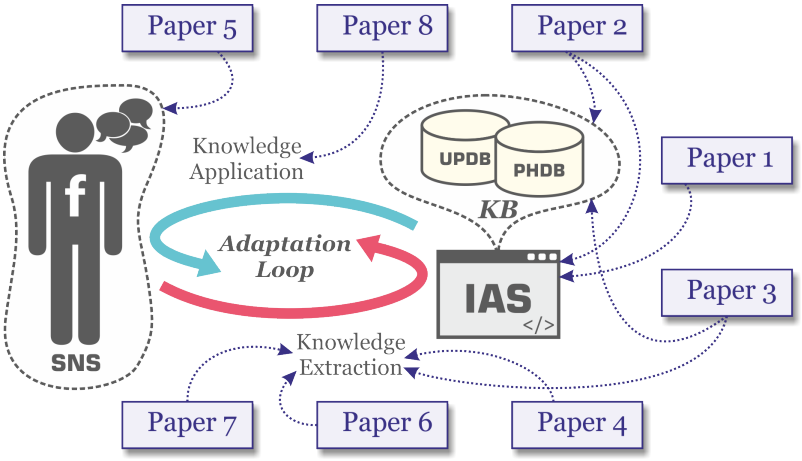


Figure 4.1: Contributions map

Fig. 4.1 illustrates the aspects of IAS that are addressed in each paper. As we mentioned, Papers 1 and 2 contribute to the development of IAS architecture (i.e. IASA). Furthermore, Paper 2 also contributes to the definition of IAS's KB (i.e. the PHDBs and UPDBs) together with Paper 3. This last one also introduces PHeDer, which is a method for *knowledge extraction* as described in section 2.3.2. Likewise, Paper 4 contributes to the process of knowledge extraction with a method for eliciting PHs out of deleted posts. Another aspect of this extraction process is treated in Papers 6 and 7 with the instantiation of the audience component of PHs (as shown in section 2.3.4). On the other hand, Paper 8 contributes to the process of *knowledge application* with an approach for estimating privacy risks and an algorithm for the generation of adaptive warning messages. Finally, Paper 5 provides an outcome which is relevant for the public sector. That is, the users of SNSs and their privacy rights.

4.2 Discussion

Privacy has been in the spotlight of research due to the high social impact of information technologies. Technology has unleashed major privacy concerns and challenges since the systems used nowadays to store and transmit personal data are not free of intrusions, leakage and other forms of cyber-attacks. Therefore, it is not surprising that a large amount of privacy research has been dedicated to matters of *security* and *control* over the dissemination of private data in information systems. Con-

[1] A report on this session can be found here: https://bit.ly/2AZKmNA

sequently, a wide range of technologies has been developed for protecting private information against unauthorized access and, ultimately, from public disclosure. In general, these aspects (i.e. security and control) remain relevant for almost every system that process and stores private information. However, in SNSs, privacy as a human *practice* acquires more importance since these are spaces in which users make their private life *public*. Hence, SNSs call for a new generation of technologies that should be oriented to support the users in their privacy decisions.

This dissertation contributes to the development of PTs which aim to guide and assist the users of SNSs in their privacy decisions. Precisely, IAS is a variant of PTs which generates personalized warning messages when users are about to disclose personal information inside their posts. As we described in section 1, PTs of such characteristics have been proposed previously [57, 10, 1]. However, these solutions often rely on a set of assumptions regarding the privacy goals of the users and their literacy level. Moreover, risk-communication strategies in SNSs like the one introduced by De et al. [10] often make assumptions about the frequency and consequence of unwanted incidents when estimating their corresponding risk values. Unlike these approaches, IAS's warning messages are generated through a personalized risk assessment and risk information collected through the users' reports. This way, interventions are tailored not only with higher accuracy, but also acknowledging the privacy goals of each particular user.

Although our approach is free of assumptions related to the estimation of risks and the individual privacy goals of the users, there are some limitations that should be taken into account. One of the most salient issues is related to the automatic recognition of private information (i.e. SAs) inside the users' posts. As we mentioned in section 2.3.1, the identification of SAs can be addressed in principle through different NLP methods such as regular expressions, named-entity recognition, or deep learning algorithms. However, users of SNSs often write their posts using sarcasm or irony [34]. This can alter significantly the meaning of a post and, hence, hinder its analysis (e.g. the sentiment of a post could be classified as negative, when in fact it is sarcastic) [27, 28]. Another issue is related to those posts with SAs that do not refer to the user herself. For instance, a post like *"Living in Stockholm may sound great...but I am sure that in winter time it must be really dark and depressing"* is expressing a negative opinion about living in Stockholm, however, it is not saying that the user who wrote it lives there. Therefore, whatever method one adopts

for identifying SAs should address these challenges, so regrettable self-disclosure scenarios can be correctly recognized.

4.3 Conclusion

Unlike the offline world, SNSs are devoid of well-evolved norms of interaction which guide socialization and self-disclosure. Therefore, the users of these media platforms often find it hard to control the scope of their actions and predict others' reactions to them. In consequence, they often engage in self-disclosure practices that they later regret. As we have discussed throughout this thesis, risk awareness plays a significant role in peoples' privacy decisions. Particularly, in the amount of private information they are are willing to disclose inside their posts. However, media technologies are often devoid of risk information that could help people understand the potential negative consequences of their disclosures. The approach introduced in this thesis contributes to the development of more privacy-aware SNSs and, consequently, to protect the users of these platforms against those privacy risks that may occur when sharing private information on the Internet.

As mentioned in section 2.1, personalization is a key aspect that must be considered when engineering user-centred technologies. In this work, we have elaborated a set of software artefacts that contribute to a more personalized support in PTs, and thereby, to a better user experience. Although these artefacts collaborate under the framework of IAS, they embody a set of design principles that can be applied when developing other PTs with similar characteristics. Such principles are *adaptivity*, *viscerality* and *supportiveness*[2]. Adaptivity refers to the ability that PTs should exhibit regarding the personalization of their interventions. That is, the ability of shaping their interventions according to the individual goals and expectations of each user. In our approach, this principle is achieved with a mechanism that regulates the content and frequency of warning messages. On the other hand, viscerality deals with the emotional link that PTs should generate between the users and their private digital data. Such a link can be promoted through a risk awareness strategy similar to the one introduced in this work. Finally, supportiveness refers to the privacy practices that PTs should recommend to the users for protecting their privacy. In our case, we have elaborated this aspect through the recommendation of personalized ACLs for narrowing the audience of the posts.

[2]These principles have been elaborated in Paper 6.

As one can observe, we have addressed the development of adaptive, visceral and supportive PTs through the orchestration of different software artefacts. From our experience, we can conclude that engineering and orchestrating these artefacts cannot be done solely from the perspective of a single discipline. In our case, we have explored a variety of techniques and methods throughout this work including Constraint-based Modelling, Decision Trees, Inductive Logic Programming, Community-detection Algorithms, and simulation models of scale-free networks. All of these techniques have contributed to the development of the software artefacts of IAS and, consequently, to the definition of a user-centred approach for risk communication in SNSs. We believe that mastering different techniques is necessary when engineering user-centred solutions. Especially, for expanding our vision regarding aspects that may largely affect the usability of the technologies we develop.

4.4 Future work

Although IAS is conceived as a PT for the protection of the users' privacy, it is ultimately a recommender system. Moreover, it is a system which requires analysing and processing personal information for shaping its recommendations. Hence, the benefits of a system like IAS come along with the privacy concerns that are characteristic of recommender systems. That is, issues related to algorithmic transparency, fairness and trust that can jeopardize the users' privacy rights. This calls for a Data Protection Impact Assessment (DPIA) of the of the different software artefacts of IAS and, consequently, the architectural modules of IASA. The notion of DPIA has been introduced in the GDPR [44] and is basically an assessment that service providers must conduct in order to identify and minimize the risks that data processing may bring to the privacy rights of data subjects (i.e. the users). Although this analysis goes beyond the scope of this dissertation, it must be taken into account when addressing a full implementation of IAS.

So far, we have investigated the role of awareness when people *share* data in SNSs. Another possible direction for further research is the application of awareness strategies when people *consume* information from media platforms. That is, the development of awareness mechanisms for assessing the trustworthiness of those news that are found on the Internet. Fuhr et al. [23] introduce the concept of Information Nutrition Labels (INLs). Similar to nutrition fact labels for food packages, INLs

provide cues about the quality of the information inside a news article. In this case, the fields contained in the INL provide Internet users with information that is relevant for judging the veracity of digital news (e.g. factuality, readability, and controversy). This approach, which is promising for dealing with fake news in social media, deserves further elaboration. Particularly, on those aspects of personalization and adaptivity that could enhance the adoption of INLs by Internet users.

Bibliography

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Computing Surveys (CSUR)*, 50(3):44, 2017.

[2] Alessandro Acquisti and Ralph Gross. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In George Danezis and Philippe Golle, editors, *Privacy Enhancing Technologies*, pages 36–58. Springer Berlin Heidelberg, 2006.

[3] Susan B. Barnes. A privacy paradox: Social networking in the United States. *First Monday*, 11(9), September 2006.

[4] Yuriy Brun, Giovanna Di Marzo Serugendo, Cristina Gacek, Holger Giese, Holger Kienle, Marin Litoiu, Hausi Müller, Mauro Pezzè, and Mary Shaw. Engineering Self-Adaptive Systems Through Feedback Loops. In Betty H. C. Cheng, Rogério de Lemos, Holger Giese, Paola Inverardi, and Jeff Magee, editors, *Software Engineering for Self-Adaptive Systems*, pages 48–70. Springer Berlin Heidelberg, 2009.

[5] Moira Burke, Cameron Marlow, and Thomas Lento. Social Network Activity and Social Well-Being. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI)*, pages 1909–1912. ACM, 2010.

[6] Gul Calikli, Mark Law, Arosha K. Bandara, Alessandra Russo, Luke Dickens, Blaine A. Price, Avelie Stuart, Mark Levine, and Bashar Nuseibeh. Privacy Dynamics: Learning Privacy Norms for Social Software. In *Proceedings of the 11th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, pages 47–56. ACM, May 2016.

[7] Emily Christofides, Amy Muise, and Serge Desmarais. Information Disclosure and Control on Facebook: Are They Two Sides of the Same Coin or Two Different Processes? *CyberPsychology & Behavior*, 12(3):341–345, 2009.

[8] Emily Christofides, Amy Muise, and Serge Desmarais. Risky Disclosures on Facebook: The Effect of Having a Bad Experience on Online Behavior. *Journal of Adolescent Research*, 27(6):714–731, 2012.

[9] Albert T. Corbett, Kenneth R. Koedinger, and John R. Anderson. Intelligent Tutoring Tystems. In M. G. Helander, T. K. Landauer, and P. V. Prabhu, editors, *Handbook of Human-Computer Interaction*, chapter 5, pages 849–874. Elsevier Science, 1997.

[10] Sourya Joyee De and Daniel Le Métayer. Privacy Risk Analysis to Enable Informed Privacy Settings. In *2018 IEEE European Symposium on Security and Privacy Workshops*, pages 95–102, April 2018.

[11] Nicolás E. Díaz Ferreyra, Tobias Hecking, H. Ulrich Hoppe, and Maritta Heisel. Access-Control Prediction in Social Network Sites: Examining the Role of Homophily. In Steffen Staab, Olessia Koltsova, and Dmitry I. Ignatov, editors, *Social Informatics*, pages 61–74. © Springer International Publishing, 2018.

[12] Nicolás E. Díaz Ferreyra, Rene Meis, and Maritta Heisel. Online Self-disclosure: From Users' Regrets to Instructional Awareness. In Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar Weippl, editors, *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 83–102. © Springer International Publishing, September 2017.

[13] Nicolás E. Díaz Ferreyra, Rene Meis, and Maritta Heisel. Should User-generated Content be a Matter of Privacy Awareness? A position paper. In Kecheng Liu, Ana Carolina Salgado, Jorge Bernardino, and Joaquim Filipe, editors, *Proceedings of the 9th International Conference On Knowledge Management and Information Sharing (KMIS 2017)*, volume 3, pages 212–216. © INSTICC/SciTePress, November 2017.

[14] Nicolás E. Díaz Ferreyra, Rene Meis, and Maritta Heisel. At Your Own Risk: Shaping Privacy Heuristics for Online Self-disclosure. In *Proceedings of the 16th Annual Conference on Privacy, Security and Trust (PST)*, pages 1–10. IEEE, August 2018.

[15] Nicolás E. Díaz Ferreyra, Rene Meis, and Maritta Heisel. Towards an ILP Approach for Learning Privacy Heuristics From Users' Regrets. In Reda Alhajj, H. Ulrich Hoppe, Tobias Hecking, Piotr Bródka, and Przemyslaw Kazienko,

editors, *Network Intelligence Meets User Centered Social Media Networks*, pages 187–197. © Springer International Publishing, August 2018.

[16] Nicolás E. Díaz Ferreyra, Rene Meis, and Maritta Heisel. Learning from Online Regrets: From Deleted Posts to Risk Awareness in Social Network Sites. Submitted for publication, 2019.

[17] Nicolas E. Díaz Ferreyra and Johanna Schäwel. Self-disclosure in Social Media: An Opportunity for Self-Adaptive Systems. In *Joint Proceedings of the 22nd International Conference on Requirements Engineering: Foundation for Software Quality (REFSQ) Co-Located Events*, volume 1564 of *CEUR Workshop Proceedings*. REFSQ, CEUR-WS.org, March 2016.

[18] Nicolás E. Díaz Ferreyra, Johanna Schäwel, Maritta Heisel, and Christian Meske. Addressing Self-disclosure in Social Media: An Instructional Awareness Approach. In *Proceedings of the 2nd ACS/IEEE International Workshop on Online Social Networks Technologies (OSNT)*, pages 1–6. ACS/IEEE, December 2016.

[19] Tobias Dienlin and Miriam J. Metzger. An Extended Privacy Calculus Model for SNSs: Analyzing self-disclosure and Self-Withdrawal in a Representative U.S. Sample. *Journal of Computer-Mediated Communication*, 21(5):368–383, 2016.

[20] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. The Benefits of Facebook "friends": Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.

[21] Silvia Facchinetti and Silvia Angela Osmetti. A Risk Index for Ordinal Variables and its Statistical Properties: A Priority of Intervention Indicator in Quality Control Framework. *Quality and Reliability Engineering International*, 34(2):265–275, 2018.

[22] Lujun Fang and Kristen LeFevre. Privacy Wizards for Social Networking Sites. In *Proceedings of the 19th International Conference on World Wide Web*, pages 351–360. ACM, 2010.

[23] Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Järvelin, Rosie Jones, Yiqun Liu, Josiane Mothe,

Wolfgang Nejdl, Isabella Peters, and Benno Stein. An Information Nutritional Label for Online Documents. *SIGIR Forum*, 51(3):46–66, 2017.

[24] Seda Gürses. *Multilateral Privacy Requirements Analysis in Online Social Networks*. PhD thesis, KU Leuven, 2010.

[25] Heikki Hiilamo, Eric Crosbie, and Stanton A Glantz. The evolution of health warning labels on cigarette packs: the role of precedents, and tobacco industry strategies to block diffusion. *Tobacco Control*, 23(1):e2–e2, 2014.

[26] IBM Group. An architectural blueprint for autonomic computing. Technical report, IBM, June 2005.

[27] Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. Automatic Sarcasm Detection: A Survey. *ACM Computing Surveys (CSUR)*, 50(5):73, 2017.

[28] Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, Mark Carman, Meghna Singh, Jaya Saraswati, and Rajita Shukla. How Challenging is Sarcasm versus Irony Classification?: A Study With a Dataset from English Literature. In *Proceedings of the 14th Australasian Language Technology Association Workshop*, pages 123–127, 2016.

[29] Jeffrey O. Kephart and David M. Chess. The Vision of Autonomic Computing. *Computer*, 36(1):41–50, January 2003.

[30] Carl Kingsford and Steven L. Salzberg. What are decision trees? *Nature Biotechnology*, 26(9):1011–1013, 2008.

[31] Spyros Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security*, 64:122–134, 2017.

[32] Airi Lampinen, Sakari Tamminen, and Antti Oulasvirta. All My People Right Here, Right Now: Management of Group Co-presence on a Social Networking Site. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, pages 281–290. ACM, 2009.

[33] Mass Soldal Lund, Bjørnar Solhaug, and Ketil Stølen. *Model-Driven Risk Analysis: The CORAS Approach*. Springer Science & Business Media, October 2010.

[34] Diana Maynard, Kalina Bontcheva, and Dominic Rout. Challenges in developing opinion mining tools for social media. *Proceedings of the@ NLP can u tag# usergeneratedcontent*, pages 15–22, 2012.

[35] Gaurav Misra and Jose M. Such. REACT: REcommending Access Control Decisions To Social Media Users. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 421–426. ACM, 2017.

[36] Gaurav Misra, Jose M. Such, and Hamed Balogun. Non-Sharing Communities? An Empirical Study of Community Detection for Access Control Decisions. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 49–56, 2016.

[37] Antonija Mitrovic. SQL-Tutor: A preliminary report. Technical report, Department of Computer Science, University of Canterbury, 1997.

[38] Antonija Mitrovic, Brent Martin, and Pramuditha Suraweera. Intelligent Tutors for All: The Constraint-Based Approach. *IEEE Intelligent Systems*, 22(4):38–45, 2007.

[39] Antonija Mitrovic and Stellan Ohlsson. Evaluation of a Constraint-Based Tutor for a Database Language. *International Journal of Artificial Intelligence in Education*, 10:238–256, 1999.

[40] Stephen Muggleton and Luc De Raedt. Inductive Logic Programming: Theory and Methods. *The Journal of Logic Programming*, 19:629–679, 1994.

[41] Melanie Nguyen, Yu Sun Bin, and Andrew Campbell. Comparing Online and Offline Self-Disclosure: A Systematic Review. *Cyberpsychology, Behavior, and Social Networking*, 15(2):103–111, 2012.

[42] Hoang-Quoc Nguyen-Son, Minh-Triet Tran, Hiroshi Yoshiura, Noboru Sonehara, and Isao Echizen. Anonymizing Personal Text Messages Posted in Online Social Networks and Detecting Disclosures of Personal Information. *IEICE Transactions on Information and Systems*, 98(1):78–88, January 2015.

[43] Melissa Pujazon-Zazik and M Jane Park. To Tweet, or Not to Tweet: Gender Differences and Potential Positive and Negative Health Outcomes of Adolescents' Social Internet Use. *American Journal of Men's Health*, 4(1):77–85, 2010.

[44] General Data Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union (OJ)*, 59:1–88, 2016.

[45] Doug Riecken. Personalized Views of Personalization. *Communications of the ACM*, 43(8):26–26, 2000.

[46] Mazeiar Salehie and Ladan Tahvildari. Self-Adaptive Software: Landscape and Research Challenges. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 4(2), May 2009.

[47] Sonam Samat and Alessandro Acquisti. Format vs. Content: The Impact of Risk and Presentation on Disclosure Decisions. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 377–384. USENIX Association, 2017.

[48] Rula Sayaf and Dave Clarke. Access Control Models for Online Social Networks. In Luca Caviglione, Mauro Coccoli, and Alessio Merlo, editors, *Social Network Engineering for Secure Web Data and Services*, pages 32–65. IGI Global, 2012.

[49] Johanna Schäwel. Paving the Way for Technical Privacy Support: A Qualitative Study on Users' Intentions to Engage in Privacy Protection. In *The 67th Annual Conference of the International Communication Association*, 2017.

[50] Petra Schubert, Leimstoll Uwe, and Daniel Risch. Personalization Beyond Recommender Systems: An Application-Oriented Overview of Personalization Functions. In Reima Suomi, Regis Cabral, J. Felix Hampe, Arto Heikkilä, Jonna Järveläinen, and Eija Koskivaara, editors, *Project E-Society: Building Bricks*, pages 126–139. Springer U.S., 2006.

[51] Luke Stark. The Emotional Context of Information Privacy. *The Information Society*, 32(1):14–27, January 2016.

[52] Charles Steinfield, Nicole B. Ellison, and Cliff Lampe. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, 29(6):434–445, 2008.

[53] Fred Stutzman, Robert Capra, and Jamila Thompson. Factors mediating disclosure in social network sites. *Computers in Human Behavior*, 27(1):590–598, 2011.

[54] Monika Taddicken. The 'Privacy Paradox' in the Social Web: The Impact of Privacy Concerns, Individual Characteristics, and the Perceived Social Relevance on Different Forms of Self-Disclosure. *Journal of Computer-Mediated Communication*, 19(2):248–273, 2014.

[55] Sonja Utz, Martin Tanis, and Ivar Vermeulen. It Is All About Being Popular: The Effects of Need for Popularity on Social Network Site Use. *Cyberpsychology, Behavior, and Social Networking*, 15(1):37–42, 2012.

[56] Sebastián Valenzuela, Namsu Park, and Kerk F. Kee. Is There Social Capital in a Social Network Site?: Facebook Use and College Students' Life Satisfaction, Trust, and Participation. *Journal of Computer-Mediated Communication*, 14(4):875–901, 2009.

[57] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. Privacy Nudges for Social Media: An Exploratory Facebook Study. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 763–770. ACM, 2013.

[58] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. "I regretted the minute I pressed share": A Qualitative Study of Regrets on Facebook. In *Proceedings of the 7th Symposium on Usable Privacy and Security, SOUPS 2011*, pages 1–16. ACM, 2011.

[59] Yi-Chia Wang, Moira Burke, and Robert Kraut. Modeling Self-Disclosure in Social Networking Sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)*, pages 74–85. ACM, February 2016.

[60] Danny Weyns, M Usman Iftikhar, Sam Malek, and Jesper Andersson. Claims and Supporting Evidence for Self-Adaptive Systems: A Literature Study. In *Proceedings of the 7th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, pages 89–98. IEEE, 2012.

[61] Zhao Yang, René Algesheimer, and Claudio J. Tessone. A Comparative Analysis of Community Detection Algorithms on Artificial Networks. *Scientific Reports*, 6(30750), 2016.

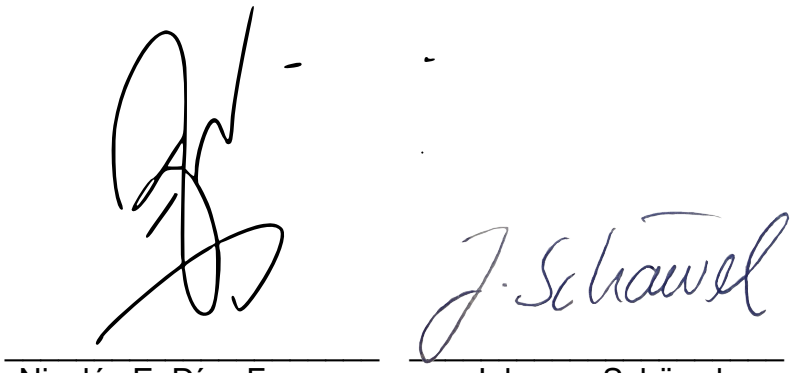
Appendix

DECLARATION OF INDIVIDUAL CONTRIBUTIONS

Publication title: Self-disclosure in Social Media: An opportunity for Self-Adaptive Systems

Reference item: N. E. Díaz Ferreyra, J. Schäwel, " Self-disclosure in Social Media: An opportunity for Self-Adaptive Systems," Joint Proceedings of the Workshops and Doctoral Symposium on Requirements Engineering - Foundation of Software Quality (REFSQ 2016).

Author	Contribution	%
Nicolás E. Díaz Ferreyra	- Conceptualisation of the approach. - Planification of the work. - Draft of the manuscript.	80%
Johanna Schäwel	- Discussion of the approach. - Draft of the manuscript.	20%



Nicolás E. Díaz Ferreyra Johanna Schäwel

Self-disclosure in Social Media: An opportunity for Self-Adaptive Systems

Nicolás Emilio Díaz Ferreyra and Johanna Schäwel

University of Duisburg Essen, Germany
{nicolas.diaz-ferreyra,johanna.schaewel}@uni-due.de
https://www.ucsm.info/

Abstract. Users of Social Network Sites (SNSs) spend considerable amounts of hours per day exchanging (consuming or sharing) information and using services provided by such platforms. However, nothing comes for free. SNSs survive at the expense of the information that users' upload to their profiles, and the knowledge derived from their on-line behavior. Discovering hidden knowledge in social networks is a centerpiece in many personalized on-line services and ad-targeting techniques, and helps to make a SNS profitable. However, users seem not to be aware of this common practice and keep sharing content compulsively. Nevertheless, self-disclosure and over-exposition can have severe consequences and can put users' integrity into risk. In order to develop better information control and awareness systems, we believe that it is important to take into account the users' on-line habits and behavior. In this work we introduce an initial assessment of the different factors that contribute to self-disclosure in Social Media, and discuss the elements that a self-adaptive solution should consider to address this issue.

Keywords: social-media, self-disclosure, awareness, self-adaptive systems

1 Introduction

Social Media has set new standards for our interpersonal relations, and has accelerated the dynamics of our lives. Many users are bridged through SNSs, and new sub-communities are built everyday based on common interests, likes or even mottos. The inhabitants of these virtual communities are spending considerable amounts of time exchanging (consuming or sharing) information, and using services provided by the SNSs. However, none of this is for free. SNSs survive at the expense of the information that users' place in their profiles, and the behavior they exhibit while using the different services provided by these platforms.

Discovering hidden knowledge in social networks is a centerpiece in many personalized on-line services and ad-targeting techniques, and is basically what makes a SNS profitable [18]. However, many of the content that is uploaded to social platforms (text, image, video, location) contain a high level of private and

sensitive information. The reality is that Social Media users compulsively share content without caring about the consequences. Moreover, their behavior off-line (in the real world) differs highly from their on-line behavior (inside a SNS) [3][14]. If we add to this that users are careless when adding new contacts to their network, there is a high chance of having potentially dangerous individuals accessing this information.

Although existing privacy-preserving mechanisms have been developed and improved over the years, they are still not helping users in distinguishing a self-exposition behavior that might put them into risk. It is very hard for a regular user to keep track of everything that he or she has shared through its "on-line life". Moreover, once the content has been shifted to the Internet, the user has no control over it anymore. This situation demands new mechanisms for tracking the sensitive information that a user has already shared, and the degree of sensitiveness that new information might have. Thus, users of SNSs can make a wiser decision before sharing content, and have a better vision of what they have shared (and would like to un-share) in the past.

In this work we present an analysis of the "self-disclosure" problem in Social Media and provide insights towards a self-adaptive solution. Different dimensions of the problem like the users' behavior and information sensitiveness are studied from an inter-disciplinary perspective. Furthermore, initial guidelines for a self-adaptive approach based on the MAPE-K model by IBM [9] are here introduced.

In the following section the fundamental bases and concepts involved in our proposal are initially introduced (Section 2). Section 3 covers the different aspects of the self-disclosure issue including: the diversity of information in SNSs, the so-called "privacy paradox", information sensitiveness, and an adapted version of the MAPE-K model. Next, Section 4 discusses alternative existing solutions, and finally Section 5 presents our conclusions and related future work.

2 Theoretical Background

This section introduces the fundamental concepts that form the bases of our proposal. Here, Autonomic Systems and run-time self-adaptation concepts are presented and analyzed for further application in a Social Media scenario.

In order to raise awareness of self-disclosure among the users of SNSs we propose to develop an Autonomic Computing vision of this issue. The goal of Autonomic Computing is to design and develop distributed and service-oriented systems that can easily adapt to changes that affect the system administration and service delivery, while reducing some of the complexities associated with the management of such systems [10]. Considering the user's content-sharing behavior in SNSs as the managed element of our autonomic system, will allow us to apply the concepts of Autonomic Computing into a Social Media domain. MAPE-K (Monitor, Analyze, Plan, Execute, and Knowledge) is a reference model for control loops used in Autonomic Computing with the objective of supporting the concepts of self-management, specifically: self-configuration, self-optimization, self-healing, and self-protection [9][10]. Fig. 1 shows the ele-

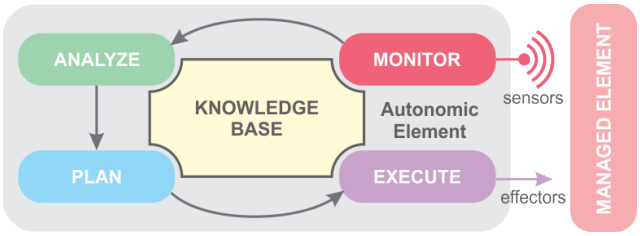


Fig. 1. Autonomic Computing and MAPE-K Loop [9]

ments of an Autonomic System: the control loop activities, sensor and effector interfaces, and the managed system.

The *Monitor* component provides the mechanisms to observe through *Sensors* different events or changes that take place in the *System* (managed element). It also filters and aggregates the data, and reports details or metrics [9]. The *Analyze* component provides the means to correlate and model the reported attributes or measurements. It is able to interpret the environment, to handle complex situations, and predict future scenarios. *Plan* provides the means to construct the set of actions required to achieve a certain goal or objective in response to certain events. On the other hand, *Execute* offers the elements to release the actions involved in a particular plan (e.g. to control the system by means of *Effectors* that modify the managed element)[10]. Additionally, a common *Knowledge Base* acts as the central part of the control loop, and is shared by the activities to store and access collected and analyzed data.

The MAPE-K model is used as an architectural reference in cases where a feedback loop is a distinctive characteristic of the system being built. Such is the case of [7], where a MAPE-K loop is used for run-time monitoring of trustworthiness properties in a socio-technical system in order to achieve trust goals. In a Social Media context like ours, the users' accounts are the elements we want to monitor since they contain the resources and services consumed by them. In line with this, the actions executed over the accounts (managed elements) are directed to aware the users about an over-exposition behavior.

3 A self-adaptive approach for addressing Self-disclosure

It is necessary to conduct an analysis of several factors that contribute to the problem and the solution of self-disclosure in SNSs. In this section we will go through the different types of information that can be found in a SNS (particularly on Facebook), and we will provide some insights for further sensitiveness classification. We will also discuss the influence of users' on-line behavior and risk aversion. At the end of this section, an approach for addressing this issue based on the MAPE-K model will be introduced.

3.1 Diversity of information in SNSs

SNSs are a rich source of the most varied kinds of information. However, users do not realize the importance that this information they "voluntarily" deposit in these sites has. From a high level inspection, normally one can find in a Facebook profile the following information: list of friends, personal information (e.g. first name, surname and profession), wall posts (public messages from other users), messages, photos, and notes [11].

However, if one takes a closer look to the now improved "Facebook Security Centre", it is now possible for users to download a copy of the information that Facebook stores about them. Surprisingly, the list is way bigger than the one mentioned before, and includes (among other information)[1]:

- *Ads Clicked:* Dates, times and titles of ads clicked by the user.
- *Ad Topics:* A list of topics that the user is targeted against based on its likes, interests and other data included in its Timeline.
- *Check-ins:* Places where the user has checked-in to.
- *Facial recognition data:* A unique number based on a comparison of the photos the user has been targeted in.
- *IP Address.*
- *Log-ins and Log-outs.*
- *Deleted friends.*

Clearly, users do not submit many of this information voluntarily to Facebook. For someone familiar within SNSs and their privacy practices, it is not surprising that Facebook (like many other SNSs) keeps all these records in their servers. However, for many users (newcomers or advanced) this situation remains unclear, even when the privacy settings of their Facebook accounts are public by default [16].

3.2 Self-disclosure and the Privacy Paradox

Exposing personal information to other persons is referred as individuals' self-disclosure. Self-disclosure in on-line contexts like Social Media is, at least to a certain extent, the precondition for a functional social network [12]. In other words, users' contributions are necessary for the survival of SNSs. Without the users' shared content (such as posted information and tagged photos), SNSs would lack of diversity and fail on being interesting enough for the users to engage with.

Self-disclosure is frequent among the users of SNSs. Furthermore, users seem careless when providing sensitive information through SNSs. However, they consider privacy protection an important issue that must be addressed. This phenomenon of contradiction has been referred as the "privacy paradox" [2][14]. Despite the studies that reveal evidence of this thesis [8], complementary research judges the non-holistic approach of the applied methods in these findings

[1] https://www.facebook.com/help/405183566203254/ (last access: 22/01/2016)

[5]. Nevertheless, we believe that, whether the privacy paradox exists or not, users' on-line behavior has to be empowered with a recommendation system that can assist them in the identification of potentially sensitive information in real time.

3.3 Defining sensitiveness in Social Media

Several gaps and dilemmas have been identified when trying to define what sensitive information is [15]. Moreover, it is a matter of discussion in the legislation of many countries and politico-economic unions [1][6]. The European Parliament for instance has defined some "personal data" categories (e.g. racial or ethnic origin) that are protected against public disclosure. It also makes use of the term "sensitive information", however it does not define it [6]. The Canadian Personal Information Protection and Electronic Documents Act 2000 state that: "Although some information (e.g. medical records) is almost always considered to be sensitive, any information can be sensitive depending on the context" [6]. This last one is an interesting approach towards the definition of "sensitive information" since it highlights the influence that the context has over it. Nevertheless, this reveals the need for considering and understanding the context where the information is placed.

A SNS is a complex environment where multiple factors converge and (in many cases) are the ones that define the rules of interaction and contributions for the users. For instance, a post that can look trivial on Facebook can be totally inappropriate in another SNS like LinkedIn (e.g. a photo of you in a party might not look very professional). In other words, here the context affects the degree of sensitiveness of the content. In this case the targeted audience of the SNS is a conditioning dimension of the context.

3.4 Towards a MAPE-K-based approach

An awareness system like the one proposed in this work has a self-adaptive nature. Its purpose is to perform a constant monitoring over the user sharing activities and notify when a self-disclosure behavior is detected. This notification can be seen as an interaction with the user, where he or she will have the last word and control over the sharing act. In other words, the user should have the chance to accept or reject the recommendation of not to share potentially sensitive information. This sequence of detection-notification-acceptance defines a feedback loop between the user and the awareness system.

As we have discussed in the previous sections, classifying information into categories of sensitiveness does not have a straightforward solution. However, as several legislations agree, it is possible to build categories of "personal" or "sensitive" data. Since the user's perception is also a determinant on the final classification, it seems logical to perform a classification of the users based on their interpretation of particular pieces of information. Then, by combining these two approaches together with attributes of the SNS (e.g. the targeted audience

Fig. 2. Elements for the analysis of sensitive information

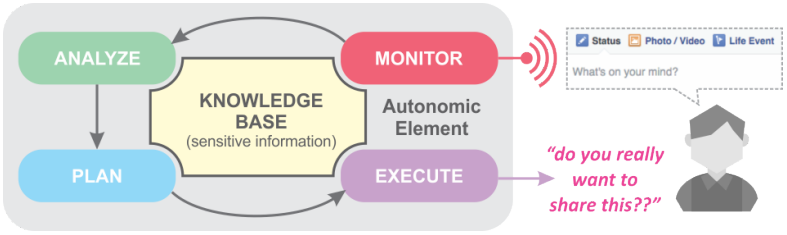


Fig. 3. Adapted MAPE-K Loop

and activity levels of the users), a better classification of the information can be performed (Fig. 2).

In Fig. 3 an adapted version of the MAPE-K loop is described. In this case, the *Managed element* corresponds to the representation of the user in a SNS, this is, the user's account. In this approach, the *Monitor* is sensing the activity of the user and responses when an information-sharing event takes place. As was previously mentioned, the goal of this system is to provide recommendations to the user when it attempts to publish content of sensitive nature. Therefore, what the *Analysis* unit should do is to analyze the information collected by the Monitor's *sensors* and classify it into sensitive or not sensitive. Here, the *Knowledge base* has a main role because it contains all what has been learned about sensitiveness and its influencing factors. After this is done, the *Plan* will elaborate a recommendation for the user, and then the *Execute* module will proceed to deliver it to the user.

A privacy protection recommendation system must be able to adapt on users individual self-disclosing behavior without destructing the interactive nature of SNSs. Our approach takes this statement into account by asking the user "do you really want to share this?" instead of forbidding it to continue. By this, the autonomy of the user is ensured and its final decision contributes to the feedback loop of the system.

Nevertheless, self-adaptation brings into account a fundamental reasoning problem: decide which is the best course of action to follow based on the perceived

stimuli from the environment. In Artificial Intelligence this type of reasoning is usually called *planning*, where the condition to achieve is called *goal* and the sequence of actions that will make the goal true is called a *plan* [4]. Because such Autonomic Element must exhibit an intelligent behavior, *planning* is a central discipline in our study. According to [4] Situation Calculus based on First Order Logic (FOL) is an adequate candidate to support *planning* due to its appropriateness for representing dynamically changing worlds. Furthermore, it provides a framework for defining a set of actions, states and changes in the environment, and entails a reasoning mechanism to make inferences. Adapting Situation Calculus to our problem domain is one of the major challenges of our research.

4 Discussion

Many privacy breaches in SNSs have been identified and addressed through different types of privacy-preserving software architectures (e.g. P2P). Many researchers advocate particularly for decentralized architecture schemas unlike predominant centralized approaches[13]. Some of the benefits of this are end-to-end encryption, hidden activity from 3rd parties, and hidden social graph among others. Although decentralized schemas improve privacy protection for the users, they demand a major development effort and cannot provide the same functionality as centralized ones [13]. This is one of the major reasons why users are reluctant to migrate to privacy-preserving SNSs [13].

While these approaches focus mainly on the architectural elements that a privacy-preserving SNS must have, the solution presented in this work propose to contribute to privacy on the application level. This is, even with a centralized and non-privacy-preserving SNS architecture, it should be possible to arise user's awareness and hence prevent extensive self-disclosure. In this way, empowered users will take better control over their on-line acts and in consequence over their private data. This can be achieved since SNSs like Facebook provide APIs and extension points for including 3rd party applications, which would allow us to integrate our solution without forcing users to change into another SNS.

5 Conclusions

In off-line situations people's communication about sensitive topics take place behind closed doors; whereas in SNSs users do not seem to lock their metaphorical doors when they address sensitive topics [3]. Moreover, the range of the audience that can access to personal information is perceived differently in on-line and off-line contexts. In an off-line context a person usually recognizes his or her audience, whereas in on-line contexts people are not able to sufficiently estimate the size of such audiences [17]. Due to the difficulty in estimating the number of receivers of what in many cases can be sensitive information, it is important to support the users in analyzing the sensitivities of their contributions.

It is true that some users are not much concerned about the consequences that self-disclosure in SNSs could bring to them, and are not willing to modify their behavior. However, this does not neglect the fact that it is necessary to support and empower them through better control and awareness systems. Instead, this raises the necessity of developing instruments that take into consideration users' distinctive characteristics that make them more or less adverse to the risks of over-exposition.

Self-disclosure and information sensitiveness analysis propose a number of challenges and opportunities for self-adaptive systems. This work has analyzed and summarized the requirements that a self-adaptive solution must cover for addressing self-disclosure in SNSs. Now this vision has to be put into practice and undoubtedly new challenges and research questions will arise. This is matter of our future work, together with an analysis of acceptance of such awareness system among the social network's community.

Acknowledgments. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

References

1. Australian Law Reform Commission, et al.: For your information: Australian privacy law and practice (alrc report 108). Sydney: Commonwealth of Australia (2008)
2. Barnes, S.B.: A privacy paradox: Social networking in the United States. First Monday 11(9) (2006)
3. Bartsch, M., Dienlin, T.: Control your facebook: An analysis of online privacy literacy. Computers in Human Behavior 56, 147–154 (2016)
4. Brachman, R.J., Levesque, H.J.: Knowledge representation and reasoning, vol. 9. Morgan Kaufmann Publishers, Massachusetts, US (2004)
5. Dienlin, T., Trepte, S.: Is the privacy paradox a relic of the past? an in-depth analysis of privacy attitudes and privacy behaviors. European Journal of Social Psychology 45(3), 285–297 (2015)
6. Directive, E.: 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the EC 23(6) (1995)
7. Gol Mohammadi, N., Bandyszak, T., Moffie, M., Chen, X., Weyer, T., Kalogiros, C., Nasser, B.I., Surridge, M.: Maintaining trustworthiness of socio-technical systems at run-time. In: Trust, Privacy, and Security in Digital Business - 11th International Conference, TrustBus 2014, Munich, Germany, September 2-3, 2014. Proceedings. pp. 1–12 (2014), http://dx.doi.org/10.1007/978-3-319-09770-1_1
8. Hughes-Roberts, T.: Privacy and social networks: Is concern a valid indicator of intention and behaviour? In: SocialCom 2013: International Conference on Social Computing. pp. 909–912. IEEE (2013)
9. Kephart, J., Kephart, J., Chess, D., Boutilier, C., Das, R., Kephart, J.O., Walsh, W.E.: An architectural blueprint for autonomic computing. IBM White paper (2003)

10. Kephart, J.O., Chess, D.M.: The vision of autonomic computing. Computer 36(1), 41–50 (2003)
11. McCown, F., Nelson, M.L.: What happens when Facebook is gone? In: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries. pp. 251–254. ACM (2009)
12. Nguyen, M., Bin, Y.S., Campbell, A.: Comparing online and offline self-disclosure: A systematic review. Cyberpsychology, Behavior, and Social Networking 15(2), 103–111 (2012)
13. Schwittmann, L., Wander, M., Boelmann, C., Weis, T.: Privacy preservation in decentralized online social networks. IEEE Internet Computing (2), 16–23 (2014)
14. Taddicken, M.: The 'privacy paradox' in the social web: The impact of privacy concerns, individual characteristics, and the perceived social relevance on different forms of self-disclosure. Journal of Computer-Mediated Communication 19(2), 248–273 (2014)
15. Thompson, E.D., Kaarst-Brown, M.L.: Sensitive information: A review and research agenda. Journal of the American Society for Information Science and Technology 56(3), 245–257 (2005)
16. Vilić, V., Radenković, I.: Privacy protection on Facebook, Twitter and LinkedIn. Syntesis: International Scientific Conference of IT and Business-Rlated Research (2015)
17. Vitak, J.: Balancing privacy concerns and impression management strategies on Facebook. In: Symposium on Usable Privacy and Security (SOUPS) (2015)
18. Zheleva, E., Terzi, E., Getoor, L.: Privacy in Social Networks. Synthesis Lectures on Data Mining and Knowledge Discovery, Morgan & Claypool Publishers (2013), `https://books.google.de/books?id=5YpiAQAAQBAJ`

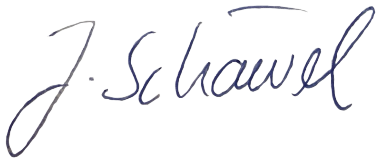
DECLARATION OF INDIVIDUAL CONTRIBUTIONS

Publication title: Addressing Self-disclosure in Social Media: An Instructional Awareness Approach

Reference item: N. E. Díaz Ferreyra, J. Schäwel, M. Heisel and C. Meske, "Addressing self-disclosure in social media: An instructional awareness approach," 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), Agadir, 2016, pp. 1-6. doi: 10.1109/AICCSA.2016.7945815

Author	Contribution	%
Nicolás E. Díaz Ferreyra	- Conceptualisation of the approach. - Planification of the work. - Draft of the manuscript.	75%
Johanna Schäwel	- Discussion of the approach. - Planification of the work. - Draft of the manuscript.	15%
Christian Meske	- Supervision and advice.	5%
Maritta Heisel	- Supervision and advice.	5%

Nicolás E. Díaz Ferreyra



Johanna Schäwel

Maritta Heisel

Christian Meske

Addressing Self-disclosure in Social Media: An Instructional Awareness Approach

Nicolás Emilio Díaz Ferreyra, Johanna Schäwel, Maritta Heisel and Christian Meske
University of Duisburg-Essen, Germany
RTG User-Centred Social Media
https://www.ucsm.info/
Email: {nicolas.diaz-ferreyra, johanna.schaewel, maritta.heisel, christian.meske}@uni-due.de

***Abstract*—Nowadays the information flowing across the different Social Network Sites (SNSs) like Facebook is highly diverse and rich in its content. It is precisely the diversity of the users' contributions to SNSs that makes these platforms attractive and interesting to engage with. Nevertheless, there is a high amount of private and sensitive information being disclosed permanently by these users in order to take full advantage of the services offered by such sites. Current privacy-protection approaches (like the one provided by Facebook) allow users to restrict the audience of their contributions and hide particular pieces of information; however, they are still far from being widely adopted and put proactively into practice. For this reason, we propose to analyze and address different aspects of online self-disclosure in Social Media from a pedagogical and self-adaptive perspective. In this work we introduce the architecture of an Instructional Awareness System (IAS) based on the MAPE-K blueprint for autonomic systems, and provide a definition of its feedback mechanism using principles of Constraint-Based Modeling (CBM).**

1. Introduction

Self-disclosure has been defined as the "process of making the *self* known to other persons" [1] and is a common activity which takes place in both offline and online contexts like SNSs. Since the rules of interaction in the online world seem to be constraint-free, users often find it difficult to self-regulate their levels of exposure when interacting through these sites. In consequence, users are more likely to reveal private and sensitive information in their contributions, leaving an open door to privacy threats (like scamming, stalking, grooming or cyber-bulling).

Research in Social and Media Psychology revealed that online self-disclosing activities are often grounded on users' personal characteristics like narcissism [2] or impression management motivations [3] [4]. For instance, users with a narcissistic personality are less engaged in using privacy features, and users who tend to engage in impression management activities do not use privacy protection tools very often [2]. Since personal characteristics of the users are unlikely to be modified, one way to mitigate the likelihood of self-disclosure harming consequences is raising the level of awareness among the users.

Up to some extent, users of SNSs can be seen as individuals who could significantly benefit from learning about online privacy management. Like tutors in academic environments (which provide guidance to students in order to help them to carry forward a particular task), an awareness mechanism that follows such premise should for instance support the users in the identification of sensitive information among their contributions as well as privacy threat scenarios. Such a pedagogical approach on self-disclosure (that considers users as learners) broadly resembles aspects of Intelligent Tutoring Systems (ITSs).

ITSs aim to engage learners (the system's users) in a sustained reasoning activity by providing guidance or feedback adapted to their learning capabilities [5]. The selection of an architecture capable to explicitly represent a control loop functionality, together with a proper knowledge representation approach, are some of the challenges when developing an ITS. In this work we address these problems jointly. In one hand, we prescribe the architectural components necesarry for the development of an Instructional Awareness System (IAS) (based on the proposal of Díaz Ferreyra and Schäwel in [6]). Simultaneously, we introduce Constraint Based Modeling (CBM) as the selected approach for representing the system's knowledge and establish the requirements for its realization.

In the following section we present the fundamental concepts that set the theoretical bases of our proposal, including ITSs and CBM. Section 3 discusses the requirements and means for the realization of an IAS for supporting online self-disclosure awareness. Such concepts are then assembled in Section 4 into an architectural approach for IASs. In section 5 a discussion of the approach in the context of existing related work is carried out. Finally, in section 6, conclusions and open research questions are summarized.

2. Theoretical Background

ITSs are computerized tutoring systems which provide their users with learning environments adapted to their knowledge and learning capabilities [7]. This adaptation process consists in selecting the right instructional content and tutorial strategy for each particular user, as well as to diagnose their knowledge state [7]. One of the main components of an ITS is a *Knowledge Base* (KB) containing

the production rules used to track how well the user is performing on the topic being taught (and consequently provide the right instructional feedback). In this section we introduce Constraint-Based Modeling (CBM), a knowledge representation approach in which feedback messages are associated with a set of constraints on correct problem solutions [8]. CBM and the reasoning mechanism that it entails will serve as the theoretical bases for the rest of this work.

2.1. Knowledge Representation in ITSs

There are at least three distinguishable approaches for knowledge representation in ITSs: propositions, procedures, and rules. Propositional representations make use of formal logic (e.g. Horn clauses) to encode the knowledge into propositions. Procedural representations include procedural and functional programming languages like Pascal and Lisp in which knowledge is represented through a hierarchy of functions (procedures and sub-procedures). On the other hand, instructional systems based on rules consist of a propositional KB, a rule set, and an interpreter that executes the rules against the KB [9].

Although each of these types of knowledge representation has strengths and weaknesses, they all share a common problem which is their over-specificity [9]. A knowledge base encoded in Horn clauses, Lisp functions or production rules requires a high level of specification and detail comprising dozens – and sometimes hundreds – of individual knowledge elements. Executing such highly articulated and detailed models demands a high amount of computational effort. Moreover, empirical data extracted from interviews is not enough to support the level of specificity these approaches require [8]. Consequently, the KB remains incomplete and lacks course-grained information about the user.

2.2. Constraint-Based Modeling

Due to the high level of specificity required by the different knowledge representation approaches, a complete and precise model of the user's knowledge is unattainable. However, even human teachers who use very loose models of their students are very effective in what they do [7]. This shows that it is possible to overcome this over-specificity problem via abstraction. Precisely, CBM proposes the use of *constraints on correct solutions* for representing the domain knowledge in order to reduce its complexity [9]. The basic assumption of CBM is that in a problem resolution process the diagnostic information about the user is not hidden in the sequence of actions he/she has performed, but in the *problem state* the user has arrived at. Consequently, users are not evaluated regarding the actions they perform, but regarding the correct state of the solutions they provide. This means that a correct solution to a problem must satisfy a set of fundamental domain principles (encoded in constraints) that cannot be violated. As long as the users never reach a state that is known to be wrong, they are free to perform whatever actions they please.

CBM is based in the *learning from performance errors* theory by Ohlsson where a constraint violation is associated with an error performed by the user [9]. Each violated constraint represents a piece of knowledge that has not been completely internalized by the user. In order to help them to incorporate the missing knowledge, users are notified when committing a violation. This early error recognition allows users to rapidly correct their actions and gain the knowledge needed for solving their tasks. Over time, the number of mistakes committed by them is reduced, and consequently their performance starts to increase [10].

3. A Constraint-Based Strategy for Addressing Self-Disclosure

So far CBM has been successfully applied for the development of ITSs in order to support learners in academic environments. Such is the case of an SQL-Tutor whose goal is to help students in formulating queries in the Structured Query Language (SQL) [7]. This time, we propose to rethink particular aspects of this strategy in order to make it applicable for a Social Media environment. In this section we treat the different challenges and characteristics that online self-disclosure demands from the IAS approach proposed in this work, including how to identify sensitive information in the users' contributions and how CBM can be used for this purpose.

3.1. Sensitive Information in Social Media

Classifying pieces of information into categories or levels of sensitiveness is a problem which raises several questions and dilemmas, especially for governments and politico-economic unions that intend to monitor and regulate the personal information stored in public or private databases [11]. For the European Parliament for instance, information regarding *racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union memberships, health or sex life* of an individual are protected against public disclosure; and it is through these categories that the concept of "sensitive information" is defined. Other approaches, like the one included in the Canadian Personal Information Protection and Electronic Documents Act 2000, states that: "Although some information (e.g. medical records) is almost always considered to be sensitive, any information can be sensitive depending on the context". This last part is particularly interesting because it highlights the influence of the context where the information is disclosed and analyzed.

We certainly believe that all information can enclose a sensitive connotation depending on the context where it is analyzed, and the information that can be found within SNSs is not an exception. For instance, certain content that may look trivial on Facebook can be very inappropriate in another SNS context like Linkedin (e.g. a photo of a person at a party). Therefore, we believe that specific categories or levels of sensitive content must be elaborated specially

for SNSs. Following, in a similar way as the European Parliament on its Directive 94/46/EC [12] or the Australian Law Reform Commission on its Privacy Act [13] we provide a list of content that can enclose sensitive information within a Social Media context.

- *Financial information* like bank account details, credit card number and financial status.
- *Biometric information*.
- *Spatial or geographical information* from check-ins in specific places and sport activities.
- *Contact information* such as address, mobile phone number, work place, etc.
- *Personal identification information* including personal ID number, passport, health security number.
- *Medical information*.
- *Political or ideological information*.
- *Sexual orientation related information*.
- *Information about religious or philosophical beliefs*.

As stated in [13], "information related to race or ethnic origin, political or religious beliefs, trade union membership and sexual orientation, is highly personal and may provide the basis for unjustified discrimination" and is therefore included in the list above. Nevertheless, these categories should not be seen as an exhaustive classification, but as an initial guideline that can be certainly improved.

Since online self-disclosure is a process were personal information is revealed to others, we propose to narrow the analysis of sensitiveness to the information that users explicitly and voluntarily submit to SNSs (namely their own posts, shared links, photos, etc.). Other information like IP address, log-ins, log-outs, and facial recognition data among others that are also kept in the SNSs servers, should not be considered as a product of self-disclosure and therefore should not be subject to analysis.

3.2. State Constraints

CBM proposes the use of *state constraints* to define the domain knowledge of an ITS and thereby overcome the over-specificity problem. As we explained in Section 2.2, one of the underlying principles of CBM is that *correct solutions* are similar to each other in the sense that they all satisfy a set of basic domain principles. No correct solution can be arrived at by traversing a problem state that violates a fundamental principle of the domain. Therefore, constraints partition the universe of possible solutions into the correct an the incorrect ones.

A state constraint is a pair of *relevance* and *satisfaction* tests on a problem state, where each member of the pair can be seen as a set of features or properties that a problem state must satisfy [8]. Therefore, the domain knowledge is configured as a collection of state constraints of the form: "if <relevance condition> is true, then <satisfaction condition> had better also be true, otherwise something has gone wrong". Constraints define classes of problem states that are pedagogically equivalent. This means that all states belonging to an equivalence class trigger the same instructional action when are violated. The relevance condition (Cr) of a constraint is the one that identifies the equivalence class for the current problem state, whereas the satisfaction condition (Cs) translates the problem state into an action if violated. In this way, feedback messages are directly attached to equivalent classes of constraints.

The targeted audience of a contribution (e.g. public, friends only, etc.) together with the categories of sensitive information proposed in Section 3.1 can be used to analyze different self-disclosure scenarios like the one illustrated in Fig. 1. In this scenario the user attempts to reveal personal contact information within a comment on a public post, which activates the satisfaction condition of a particular constraint. However, because of having included information that can be cataloged as sensitive, the satisfaction condition of the constraint gets violated. Consequently an instructional action encoded as a warning message is send to the user.

Problem State: The user writes a comment on a friend's public post. In the comment the user includes its phone number, time and place for a meeting.

Constraint 1

Cr: *The user writes a comment containing personal contact information.*
Cs: *The post's audience must be anything but public.*

Figure 1. A constraint of a comment on a public post

To some extent, *state constraints* resemble many aspects of the *production rules* used for the realization of expert systems. The difference lies in that constraints are pieces of *evaluative knowledge*, whereas a rule is a piece of *generative knowledge*. This means that whereas the first one is an instrument for passing judgment, the last one is used for computing new results or interfering new conclusions. In any case, both sides of a state constraint (Cr and Cs) can be represented as *patterns*, which are sets of conditions that are understood conjunctively (as in the antecedent side of a production rule). Therefore state constraints like production rules can be for instance specified as pairs of complex functional predicates.

4. An Instructional Awareness Approach

An IAS requires a reasoning mechanism for selecting personalized feedback messages based on an evaluation of the user's performance. In this section we introduce an approach to evaluate the progress of each particular user in the process of learning online privacy preserving principles. Since such approach requires the support of a software architecture, we have developed an architecture describing the basic components that an IAS for addressing self-disclosure should include. Both approaches, user's evaluation and architecture, are presented and discussed in this section.

4.1. The IAS User Model

In order to provide the right instructional feedback, an IAS like the one here described must measure in some way the progress of its users. ITSs that use CBM to encode their knowledge bases keep a User Model for this purpose. This model basically consists in the history of relevant and violated state constraints associated with the different problem states that a user has arrived to. For instance, the problem state in Fig. 1, is clearly violating Constraint 1 because it does not fulfill its satisfaction condition. If we slightly change the problem state to "The user writes a comment on a friend's *private* post..." and compare it once again against Constraint 1, we can see that such constraint is not being violated. However, the problem state fulfills its relevance condition and therefore the constraint is relevant. In other words, not every constraint in the KB is relevant for every problem state, and not every relevant constraint is always violated.

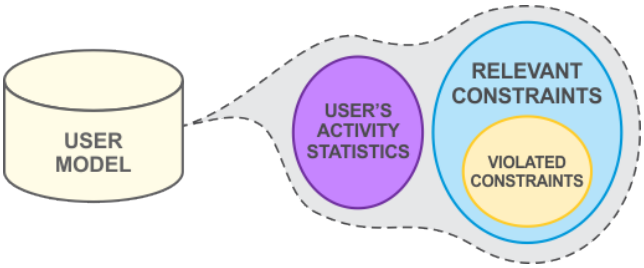


Figure 2. Elements of the IAS User Model

This distinction between relevant and violated constraints is necessary in order to properly model the user's knowledge and performance. Users that often violate state constraints in their problem states require closer feedback than the ones who don't. However, as self-disclosure activities are grounded in personal characteristics of the users, it is very likely that users with high levels of narcissism or risk-aversion will discard most of the warning feedback. This scenario needs to be contemplated by an IAS, therefore additional statistical information related to the user's activity level must be considered in order to for instance adapt the frequency and content of the feedback. Such statistical information can include the number of ignored warnings and number of interactions (posts, comments, etc.) within the SNSs among others.

4.2. The IAS Architecture

The generation of feedback based on the evaluation of the IAS User Model requires the support of a software architecture capable to provide explicit control loop functionality. For this purpose we have developed an Instructional Awareness Software Architecture (IASA) based on the MAPE-K blueprint for autonomic computing [14]. Such approach is illustrated in Fig. 3 where the MAPE-K modules (*Monitor*, *Analyze*, *Plan*, *Execute*, and *Knowledge Base*) are specified to fulfill the IAS requirements. Here, the feedback loop is characterized by the reasoning mechanism entailed by the knowledge representation approach, which in this case is CBM.

4.2.1. The Inputs. The inputs to the IAS represented in Fig. 3 are the contributions of the user within a particular SNS (status updates and comments) and are basically encoded in the form of plain text, image, video or other types of media format. The output, on the other hand, is a personalized self-disclosure warning containing insights about the state constraint that the user is violating and hints about how this can be amended. Both, contribution and feedback, are the units of information that the IAS and the environment exchange through the Sensors and Effectors connected to the *Monitor* and *Execute* modules respectively.

4.2.2. The Environment Interface. Basically, the stimuli coming from the environment is perceived as an *event* that the system must react to. This means that the IAS must perform an *action* (a function) that is opportunely invoked to *handle* that particular event. Since this interaction between the IAS and the environment (the SNS) is performed by the *Monitor* and *Execute* modules, they are both endowed with Listeners and Event Handlers to support this task. We have assumed that the communication between the SNS and the IAS is done though an API which varies from platform to platform (e.g. the API from Facebook is different to the one from Twitter). Therefore, and in order to lower the coupling between the environment and the IAS, we have grouped both *Monitor* and *Execute* modules under a common Environment Interface (EI).

4.2.3. Information Analysis. Many irrelevant and noisy data can be present in the input that can hinder its analysis. Consequently, this can impact negatively in the quality of the output generating inconsistencies in the feedback. This requires a data conditioning process previous to the analysis that is carried out by a Pre-Process component inside the *Analyze* module. The interface M-I is provided by this component to the EI in order to forward the input from the environment to the system. As it is shown in the IAS User Model, there is information related to the user's *activity statistics* (e.g. sharing frequency, audience, number of friends, discarded warnings) that can be retrieved from the environment and can certainly contribute as adaptability variables of the IAS. Therefore the Pre-Process component must also disaggregate such information from the input. Consequently, two data sets are obtained after pre-processing the input: data regarding the *user's activity statistics* and the data concerning the contribution (in other words the *problem state*). The first one is used to compute statistical values related to the online activity of the user and is later stored in the KB as part of the User Model (via the interface UM-I1). The second is further compared against the State Constraints to pinpoint self-disclosure patterns in the user's contribution. Such task is carried forward by the Constraint Match component which together with the

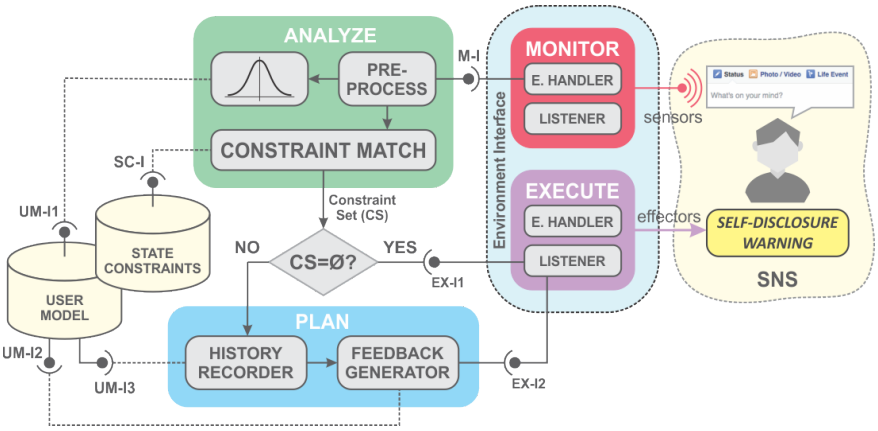


Figure 3. The IAS Architecture (IASA)

Statistical, Pre-Process and Constraint Match components comprise the *Analyze* module.

4.2.4. Constraint Matching. One of the advantages of CBM over other knowledge representation approaches is that it reduces the problem state diagnosis to pattern matching. Both, relevance and satisfaction condition of a state constraint, are combinations of patters that can be matched against the problem state. This task is simple but can certainly consume considerable amounts of computational effort and time, specially when the number of patterns is large. This problem can be addressed by representing the state conditions (Cr and Cs) in compiled forms like RETE networks [7]. The advantage of such representation is that it allows applying the conditions that appear in many patterns only once, avoiding to perform a complete re-evaluation of the domain constraints in every cycle of the feedback loop. The outcome of the Constraint Match module is a Constraint Set (CS in Fig. 3) containing the constraints that are relevant for the current problem state.

4.2.5. Feedback Generation. After CS is computed, it is forwarded to the *Plan* module consisting of a History Recorder component and a Feedback Generator. If CS is empty (contains no relevant state constraints) then the control is delegated to the EI (more precisely to the Execute module) though the interface EX-I1 and produces no feedback. Otherwise the subgroup of violated constraints is computed in the History Recorder and together with CS is sent to the User Model through the interface UM-I3 to quantify the progress of the user. The Feedback Generator on the other hand produces the appropriate instructional actions according to the information stored in the User Model; that is, the history of relevant and violated constraints together with the statistical information on the user's online activity. The information contained in the User Model should be used to estimate the risk aversion of the user and consequently determinate the feedback generation frequency and the amount of information to be provided (e.g. full or partial feedback, number of hints, etc.). The interface UM-I2 is used to retrieve such information, and once the feedback is produced it is sent to the Execute module of the EI through the interface EX-I.

5. Discussion on Related Work

Average users are not completely familiar with the privacy policies of SNSs and with the available mechanisms that allow to regulate the extent of their audiences. Newcomers in the use of these social media platforms are probably the most disadvantaged ones since they hold limited knowledge about such privacy preserving instruments. Moreover, privacy preferences sometimes introduce many options that can certainly overwhelm and generate confusion over the users and consequently may impair their proactive intentions. This situation has encouraged many researchers to venture in the area of recommender systems for privacy settings in order to provide automatically generated privacy policies to cover the individual needs and expectations of the users [15].

One of the approaches that have been investigated for the automatic generation of privacy policies is Machine Learning (ML). Such is the case of [15] where the authors examine the attitude of the users when sharing photo albums on Facebook as an indicator of their risk aversion. Such indicator is later used to classify them into fundamentalist (highly cautious about revealing their personal information), pragmatist (with average privacy concerns) and unconcerned in order to reveal implicit relations between a user's attitudes towards privacy and its personal characteristics and interests. Thereafter, the privacy policies that users within a cluster have put into practice can be recommended to other users with similar characteristics.

A similar approach described in [16], introduces a machine learning privacy wizard for social media data based on implicit relations between the existing user's privacy settings and the community structure of the underlying social network. In this work the authors observed that the neighborhood network of a user can be subdivided in clusters of users (communities) that share common privacy restrictions.

This basically suggests that users build their own privacy preferences following an implicit set of rules related to the community structure of their network of friends. By analyzing the structures of such communities, the authors have developed a system capable to automatically recommend detailed privacy settings with minimal intervention from the user.

From our perspective, the development of a proactive privacy behavior is an ongoing learning process where the participation of the user is central. Approaches like the ones mentioned above mainly focus on the generation of a privacy policy, and put no (or very little) emphasis on the pedagogical aspect of the problem. The work presented in this paper aims to provide the basis (namely architecture, knowledge representation and reasoning mechanism) for the realization of a solution which explicitly cover this aspect. The knowledge representation and the reasoning mechanism of CBM supported by an architecture like MAPE-K meet in IASA for serving this purpose and addressing the evolving privacy requirements of each particular user.

One precondition for a successful guided discovery learning is the acquisition of declarative knowledge by the learners [7]. ITSs can assume that users have acquired such knowledge during lectures basically because they are designed to operate on an academic application domain. In a Social Media environment this precondition is absent since normally users are not taught how to use these platforms and therefore might not be familiar with their privacy settings. In that sense our approach plays the role of a mentor whose feedback messages should be carefully designed to overcome this gap.

6. Conclusion

As we have pointed out, users of SNSs might find it hard to self-monitor and regulate their disclosing behavior, and to some extent ignore the potential risks and harms of such practice. Recommender privacy systems address this issue by automatically generating a privacy policy for the user allowing to abstract them from the sometimes tedious work of configuring its own privacy settings. However, such approaches do not help users in reasoning holistically about their behavior and privacy practices, instead they provide a set of privacy preferences based on the current or past behavior of the user.

In this work we have presented an approach for addressing a privacy related problem like it is online self-disclosure, but from a pedagogical perspective. As previously mentioned, we believe that the constructs that influence the extent of online self-disclosure are very unlikely to be modified. Therefore we suggest to raise the awareness levels of the users as a viable way of addressing this problem. Nevertheless, it is necessary to contribute interdisciplinary in this research in order to provide the right awareness and control mechanisms that can cover the expectations of a diverse group of users. We expect this work will serve as a common ground for different scientific disciplines to collaborate and conduct further research.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

References

[1] S. M. Jourard and P. Lasakow, "Some factors in self-disclosure." *The Journal of Abnormal and Social Psychology*, vol. 56, no. 1, p. 91, 1958.

[2] S. Utz and N. Krämer, "The privacy paradox on social network sites revisited: The role of individual characteristics and group norms," *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, vol. 3, no. 2, 2009.

[3] N. Krämer and N. Haferkamp, "Online self-presentation: Balancing privacy concerns and impression construction on social networking sites," in *Privacy Online*. Springer, 2011, pp. 127–141.

[4] J. Vitak, "Balancing privacy concerns and impression management strategies on Facebook," in *Symposium on Usable Privacy and Security (SOUPS)*, 2015.

[5] A. T. Corbett, K. R. Koedinger, and J. R. Anderson, "Intelligent Tutoring Tystems," *Handbook of Human-Computer Interaction*, pp. 849–874, 1997.

[6] N. E. Díaz Ferreyra and J. Schäwel, "Self-disclosure in Social Media: An opportunity for Self-Adaptive Systems," in *Joint Proceedings of the Workshops and Doctoral Symposium on Requirements Engineering*. REFSQ, March 2016.

[7] A. Mitrovic, "SQL-Tutor: A preliminary report," Department of Computer Science, University of Canterbury, Tech. Rep., 1997.

[8] A. Mitrovic and S. Ohlsson, "Evaluation of a constraint-based tutor for a database language," *International Journal of Artificial Intelligence in Education*, vol. 10, pp. 238–256, 1999.

[9] ——, "Constraint-based knowledge representation for individualized instruction," *Computer Science and Information Systems (ComSIS) Journal*, vol. 13, no. 1, 2006.

[10] A. Mitrovic, K. R. Koedinger, and B. Martin, "A comparative analysis of cognitive tutoring and constraint-based modeling," in *User Modeling 2003*. Springer, 2003, pp. 313–322.

[11] E. D. Thompson and M. L. Kaarst-Brown, "Sensitive information: A review and research agenda," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 3, pp. 245–257, 2005.

[12] E. Directive, "95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," *Official Journal of the EC*, vol. 23, no. 6, 1995.

[13] Australian Law Reform Commission *et al.*, "For your information: Australian privacy law and practice (alrc report 108). Sydney: Commonwealth of Australia," 2008.

[14] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003.

[15] K. Ghazinour, S. Matwin, and M. Sokolova, "YOURPRIVACYPROTECTOR, A recommender system for privacy settings in social networks," *International Journal of Security, Privacy and Trust Management (IJSPTM)*, vol. 2, no. 4, August 2013.

[16] L. Fang, H. Kim, K. LeFevre, and A. Tami, "A privacy recommendation wizard for users of social networking sites," in *Proceedings of the 17th ACM conference on Computer and communications security*. ACM, 2010, pp. 630–632.

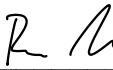
DECLARATION OF INDIVIDUAL CONTRIBUTIONS

Publication title: Online Self-disclosure: From Users' Regrets to Instructional Awareness

Reference item: N. E. Díaz Ferreyra, R. Meis, and M. Heisel, "Online Self-disclosure: From Users' Regrets to Instructional Awareness," in *1st International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE)*. Springer LNCS, September 2017, pp. 83–102. doi: 10.1007/978-3-319-66808-6_7

Author	Contribution	%
Nicolás E. Díaz Ferreyra	- Conceptualisation of the approach. - Planification of the work. - Draft of the manuscript.	75%
Rene Meis	- Discussion of the approach. - Planification of the work	20%
Maritta Heisel	- Supervision and advice.	5%

Nicolás E. Díaz Ferreyra



Rene Meis

Maritta Heisel

Online Self-disclosure: From Users' Regrets to Instructional Awareness

Nicolás Emilio Díaz Ferreyra, Rene Meis, and Maritta Heisel

University of Duisburg Essen, Germany
{nicolas.diaz-ferreyra,rene.meis,maritta.heisel}@uni-due.de
https://www.ucsm.info/

Abstract. Unlike the offline world, the online world is devoid of well-evolved norms of interaction which guide socialization and self-disclosure. Therefore, it is difficult for members of online communities like Social Network Sites (SNSs) to control the scope of their actions and predict others' reactions to them. Consequently users might not always anticipate the consequences of their online activities and often engage in actions they later regret. Regrettable and negative self-disclosure experiences can be considered as rich sources of privacy heuristics and a valuable input for the development of privacy awareness mechanisms. In this work, we introduce a Privacy Heuristics Derivation Method (PHeDer) to encode regrettable self-disclosure experiences into privacy best practices. Since information about the impact and the frequency of unwanted incidents (such as job loss, identity theft or bad image) can be used to raise users' awareness, this method (and its conceptual model) puts special focus on the risks of online self-disclosure. At the end of this work, we provide assessment on how the outcome of the method can be used in the context of an adaptive awareness system for generating tailored feedback and support.

Keywords: social network sites, adaptive privacy, awareness, heuristics, risk analysis

1 Introduction

Nowadays, different SNSs support a wide and diverse range of interests and practices [4]. While sites like Facebook or Twitter serve as more general purpose platforms, others like LinkedIn or Researchgate provide a more specific structure designed for targeting the needs of particular groups of users (professionals and scientists, respectively) [15]. Independently of their aim, the anatomy of any SNS consists of a set of core features that allow users to share, co-create, discuss and modify different types of media content [15]. Through such features users share their interests, emotions, opinions and beliefs with a large network of friends and acquaintances within a few seconds.

The act of revealing personal information to others is commonly known as "self-disclosure" [2]. This practice (which is common and frequent in both online

and offline contexts) is key for the development and maintenance of personal relationships [31]. However, disclosures (specially in online contexts like SNSs) very often reveal detailed information about the user's real life and social relationships [14]. Furthermore, when revealing too much personal information users take the risk of becoming victims of privacy threats like stalking, scamming, grooming or cyber-bulling. These threats, together with negative consequences for the user's image, make online self-disclosure in many cases a regrettable experience.

There are diverse factors which contribute to engaging in online self-disclosure activities. A poor understanding of the size and composition of audiences, psychological factors like narcissism [27] and impression management [16][28], or low privacy literacy [26] are often discussed and analyzed as the main factors mediating in online self-disclosure. However, the role of computers as social actors and consequently the role of technology in shaping our perceptions of information privacy is often omitted [25]. Since private digital data is intangible and only perceived through the interfaces and physical materials of media technologies, such technologies modulate users' emotional perception and attachment towards their private information [25]. Nevertheless, media technologies are not succeeding in taking such emotional perception to the *visceral* level. This is, making the tie between users' feelings and data visible, tangible and emotionally appreciable so they can perceive (in a visceral way) the impact of their disclosures.

Since regrettable online self-disclosure experiences often come along with a *visceral reaction*[1], they can be considered as sources of privacy heuristics which can help the users in making better and more informed privacy decisions, as to contribute in the emotional attachment towards their digital data. Díaz Ferreyra et al. [8] propose an Instructional Awareness Software Architecture (IASA) that prescribes the components of an adaptive Instructional Awareness System (IAS), which provides tailored feedback on users' disclosures in SNSs. In line with this approach, this work proposes to encode the outcome of empirical research and everyday online self-disclosure experiences into the knowledge base of IAS. Taking regrettable user experiences as the starting point, this work introduces a method for the derivation of privacy heuristics (best practices) and their further incorporation into IAS.

The rest of the paper is organized as follows. In the next section we discuss preventative technologies in the landscape of privacy technologies. In Section 3 we discuss how empirical research on users' regrettable disclosures can be a rich source of privacy heuristics and serve for the development of preventative technologies. Next, Section 4 introduces the conceptual model and the method's steps for the derivation of privacy heuristics. In Section 5, we provide assessment towards the evaluation of the method and its outcome for the generation of instructional awareness. We next discuss the advantages and drawbacks of this approach together with future work in Section 6. Finally, we conclude with an outline of the implications of our approach.

[1] A visceral reaction is an "instinctive, gut-deep bodily response to a stimulus or experience"[1]. For instance, a burning sensation in the stomach when loosing something of value (e.g. wallet, passport, etc.)

2 Related Work

Whether in or out of the context of SNSs, privacy is certainly a multifaceted and complex problem that receives the attention of researchers across a wide spectrum of disciplines. Online self-disclosure and its unwanted consequences have been discussed and treated by research in media psychology and computer science, among others. However, framing self-disclosure as a privacy problem may sound paradoxical since this is a voluntary act performed by the users, and it does not violate "the right of the individual to decide what information about himself should be communicated to others and under what circumstances" (which is Westin's definition of privacy [32]). Nevertheless, the privacy landscape is much broader and existing solutions rely on different technical and social assumptions as well as definitions of privacy [7].

2.1 Self-disclosure in the Privacy Landscape

Gürses and Díaz [7] describe the landscape of privacy technologies in terms of three paradigms: control, confidentiality and practice. Technologies located in the "control" paradigm understand privacy as Westin does (i.e. the ability to determine acceptable data collection and usage) and seek to provide individuals with control and oversight over the collection, processing and use of their data. In the "confidentiality" paradigm, technologies are inspired by the definition of privacy as "the right to be alone" and aim to create an individual autonomous sphere free from intrusions. Both paradigms, control and confidentiality, have a strong security focus but do not put much attention on improving transparency and enabling identity construction [7]. After all, privacy contributes widely to the construction of one's identity both at an individual and collective level. That is precisely the (implicit) notion of privacy that users put into "practice" when they self-disclose, namely "the freedom of unreasonable constraints on the construction of one's own identity". In order to support the users in building such constraints, technologies in the practice paradigm aim to make information flows more transparent through feedback and awareness [7].

2.2 Preventative Technologies

Many efforts have been put in raising privacy awareness among the users of SNSs in order to mitigate the unwanted consequences of online self-disclosure [6][8][9][11][29]. However, many of these preventative technologies rely on static and non adaptive awareness solutions, which in many cases hinders the engagement of the users towards such systems. Wang et al. [29] developed three plugins for Facebook which aimed to help the users to avoid regrettable disclosures. These plugins called "privacy nudges" intervened when the user was about to post a message in his/her biography either (i) introducing a delay, (ii) providing visual cues about the audience of the post, or (iii) giving feedback about the meaning (positive or negative) of the post. Despite its novelty, mixed reactions were observed when these nudges were tested against Facebook users: some users

liked them and managed to engage with them, and some others did not. An explanation to this can be found in a qualitative study conducted by Schäwel and Krämer [23], which revealed that the engagement level of privacy awareness systems is tightly related with their ability of providing tailored feedback to the users.

To overcome the issues of static approaches, other preventative technologies focus on providing personalized feedback and guidance to the users though adaptive mechanisms. For instance, Caliki et. al. developed "Privacy Dynamics", an adaptive architecture which uses Social Identity Theory (SIT) to learn privacy norms from the users' sharing behaviors [6]. Basically, the SIT postulates that people belong to multiple social identities. For instance, being *Sweedish*, being an *athlete*, or being a *researcher* are all examples of social identities/identity groups. Social identities and identity groups play an important role in the construction of people's privacy because they are tightly related to the targeted audience of the user's disclosures. This is, a user frequently has a mental conceptualization of the different social identity groups with whom he/she is interacting. However, there can be a misalignment between this mental model and the real audience, which can lead to a privacy violation. For instance, when disclosing a negative comment about one's workplace without thinking that a work colleague can be part of the post's audience. In this case the conceptualized audience is not including the work colleagues, while the actual audience is. To overcome this issue, "Privacy Dynamics" uses Inductive Logic Programming (ILP) to learn these privacy rules and consequently resolve the conflicts among them. Other adaptive solutions like the ones from Ghazinour et al. [11], and Fang et al. [9] follow similar supervised learning approaches. This work provides an instrument for the incorporation of user-centered privacy requirements into the design process of adaptive preventative technologies.

3 Theoretical Background

Regrettable online self-disclosure experiences are hardly taken into consideration for the development of preventative technologies. In this section we discuss the importance of such experiences for eliciting user-centered privacy requirements as for the generation of adaptive feedback and awareness. Likewise, we will discuss the role of regrets in the derivation of privacy heuristics and their incorporation into the design of preventative technologies.

3.1 Self-disclosure Privacy Concerns

Systems are developed on the basis of requirements that specify their desired behavior in a given environment. Privacy requirements represent the positions and judgments of multiple stakeholders with respect to privacy and transparency claims in a system-to-be [13]. In order to discuss privacy claims from a multiple stakeholders perspective, all the information that will be collected, used, processed, distributed or deleted by the system-to-be should be deemed relevant

for privacy analysis [13]. Typically, in a requirements elicitation process, stakeholders are the ones who put the privacy claims on the table for their consideration and later realization into privacy preserving features of the system-to-be. However, online self-disclosure begins when the system is up-and-running and operated by its users. Thus, privacy requirements that arise as consequence of online self-disclosure activities are mostly manifested in the operating stage of the system-to-be. Moreover, the origin of a online self-disclosure privacy concern is often a regrettable experience encountered by the user or his/her inner circle of friends, family or acquaintances.

3.2 Regrets in SNSs

Basically a regret can be defined as an unwanted consequence (factual or potential) of an action which materializes an unwanted incident (such as stalking, identity theft, harassment, or reputation damage) and derives in a feeling of sadness, repentance or disappointment [30]. Wang et al. [30] conducted an empirical study over 321 active Facebook users in order to identify different regrettable scenarios. Such regrets were identified through online surveys and interviews where users answered the question "Have you posted something on Facebook and then regretted doing it? If so, what happened?". Users reported situations where posting about (a) alcohol and illegal drug use (b) sex (c) religion and politics (d) profanity and obscenity (e) personal and family issues (f) work and company and (g) content with strong sentiment, had lead them to negative online experiences. This suggests that online self-disclosure privacy requirements do not emerge as a concern per-se, but as a consequence of regrettable online activities. Therefore, the first step into a user-centered privacy analysis should be to consider regrettable self-disclosure experiences as explicit manifestations of privacy concerns.

3.3 Instructional Awareness

In line with the adaptive preventative technologies, Díaz Ferreyra et. al. introduced the concept of IAS which consists in providing adaptive privacy guidance to the users when they intend to reveal private and sensitive information in a post [8]. IAS has its basis in IASA, which resembles principles of self-adaptation in order to satisfy the particular privacy concerns of the users. In order to provide personalized privacy guidance and feedback to the user, IASA senses the user's "post" events and identifies pieces of private and sensitive information contained in such messages. If information of such nature is indeed detected by IAS, the system proceeds to the generation of personalized feedback to inform the user about this situation. Such feedback consists in a warning message together with a recommendation about the possible preventive actions that the user can follow in order to protect his/her privacy. For example, if the user attempts to disclose his/her new phone number in a post, IAS will raise a warning message like "Your phone number is included in the post. Do you want to know how

to protect your private data?" and recommend the user to restrict the post's audience (for instance to "friends only").

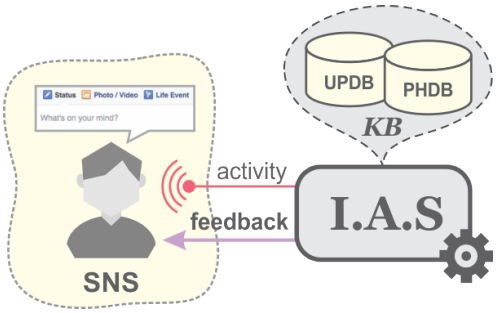


Fig. 1. Instructional Awareness System (IAS)

As shown in Fig. 1, IAS uses a Knowledge Base (KB) which is divided in two for the generation of adaptive feedback. The first one is a User Performance Data Base (UPDB) which tracks the privacy practices of the user's towards the recommendations delivered by IAS. This is, how many times the user has ignored/accepted the system's warnings, and how often the user discloses private and sensitive information, among other variables of adaptation. Such adaptation variables allow IAS to regulate the frequency and intensity of the feedback. The second part of the KB is a Privacy Heuristics Data Base (PHDB) which stores privacy knowledge encoded into constraints. Such constraints are privacy best practices which are evaluated when a "post" action takes place. Following the phone number example, if a constraint defined as "*if* post contains phone number *then* keep the audience not public" is violated, then IAS raises a warning message. As described, the UPDB and PHDB work closely together in detecting risky disclosures and recommending preventive actions to the user. In order to embody the design of IAS with user-centered privacy requirements, we propose to incorporate knowledge about online self-disclosure regrettable experiences inside the PHDB. This work will focus on the derivation of such knowledge in the form of privacy heuristics and their incorporation as the core components of IAS's PHDB.

4 Privacy Heuristics Derivation (PHeDer)

In this section we introduce the conceptual model for conducting self-disclosure privacy analysis, and our method for extracting of privacy heuristics from the users' regrettable online self-disclosure experiences. The method, called Privacy Heuristics Derivation method (PHeDer), starts with the identification of a regrettable scenario and concludes with one or more privacy heuristics defined as constraints for their later inclusion into IAS's PHDB.

4.1 Conceptual Model

In a traditional requirements engineering approach, a concern is basically raised due to actions performed over a piece of information that can lead to a privacy breach. Such actions, that when performed materialize a risk, are defined as privacy threats. The case of online self-disclosure has the particularity that the threat which exposes the user to a privacy risk is an action performed by the user him/herself. This is, the act of disclosing private or sensitive information in a post within a SNS. Thus, awareness mechanisms would enrich their performance by incorporating in their feedback engine the knowledge abut the risks of online-self disclosure. Consequently, by being informed about the possible risks of online self-disclosure, users can make more informed and wise decisions in order to protect their private information against their own actions.

The conceptual elements that form the basis for the analysis of self-disclosure experiences are represented in the Unified Modeling Language (UML) [12] class diagram of Fig. 2. As said, *Threats* are *Actions* performed over pieces of *Surveillance Information* (SI) (see Section 4.1) in the system which can lead to an *Unwanted Incident* (such as identity theft, harassment, or reputation damage). A *Post* in a SNS is a type SI which is disclosed to a specific *Audience* and is composed by one or more *Surveillance Attributes* (SA) (see Section 4.1). As mentioned, *Information Disclosure* is the *Threat* of which we want to protect the user in order to avoid a regrettable online experience. Hence, the *Absence of Regret* is the *Asset* that must be protected. A *Regret* can be factual or potential in the sense that can be the result of concrete user experiences, or the result of conceptual (not yet reported by the users) self-disclosure scenarios.

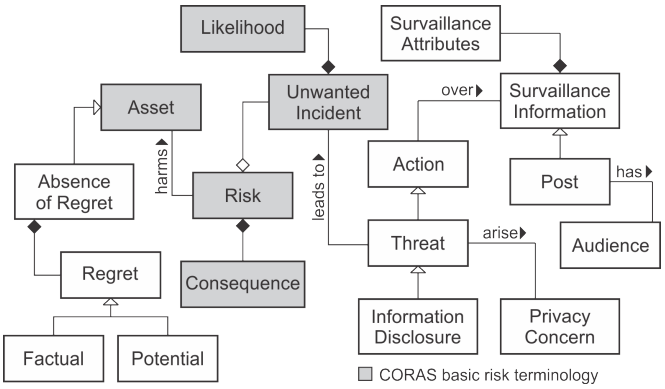


Fig. 2. PHeDer conceptual model

The PHeDer conceptual model is based on the CORAS basic risk terminology [17]. Like in CORAS, a *Risk* in PHeDer is the *Likelihood* of an *Unwanted Incident* and its *Consequence* for the *Asset*. In this sense, a *Consequence* is a value on an impact scale such as *insignificant*, *minor*, *moderate*, *major* or *catastrophic*. Likewise, the *Likelihood* is a value on a frequency scale such as *rare*,

unlikely, *possible*, *likely* and *certain*. CORAS considers that different *Assets* can be harmed by the same *Unwanted Incident* and cause different *Consequences*. Therefore CORAS models the relation between *Unwanted Incidents*, *Assets* and *Consequences* as a *Risk*. Since in our case, the only *Asset* that should be protected is the *Absence of Regret*, we will concentrate our analysis on the *Unwanted Incidents* and consider the *Risks* as such.

Risks Performing a detailed risk analysis of online self-disclosure goes beyond the scope of this work, but certainly risks must be taken into consideration when describing a self disclosure scenario. Petronio [22] describes the most common types of self disclosure risks and groups them into five categories:

- *Security risks* are situations of disruption of power that jeopardize the safety of the user or its inner circle of friends and family. For instance, a mother may be careful on revealing that her underage daughter is pregnant for fear of negative repercussions. Likewise, individuals with HIV often keep their health status information private based on the perceived safety risks (e.g. harassment, job loss, etc.).
- *Stigma risks* are grounded in the individual's self-identity and involve information that has the potential to discredit a person. These risks are based on the assumption that others might negatively evaluate individuals' behaviors or opinions. For instance, sharing controversial opinions or thoughts (e.g. religious beliefs, political affiliation, etc.), can lead to negative evaluation and even exclusion from a group.
- *Face risks (self-image)* are associated with a potential feeling of embarrassment or loss of self-image. Therefore, these situations comprise the user's internal moral face (shame, integrity, debasement, and honor) and his/her external social face (social recognition, position, authority, influence and power). For example, revealing failing in a driving test can be embarrassing.
- *Relational risks* represent situations where the disclosure of a thought or opinion might threaten the status of a relationship. Relational risks may come in a variety of forms like hurting another person's feelings by expressing negative opinions towards him/her, or expressing the concern to a partner that he/she is having an affair.
- *Role risks* take place when the disclosure of intimate information jeopardizes the social role of an individual. These are situations where the revelation of private information is perceived as highly inappropriate by the receptors. For instance, a supervisor's leader role might be compromised if he/she asks for an advice regarding his/her marital status to a subordinate.

According to Petronio [22], the risk levels of self-disclosure episodes vary from individual to individual. This is, episodes that might be seen as highly risky for some users, may not be seen as such by others. In consequence, the risk levels of self-disclosure fluctuate along a range of values in a risk scale [22]. A risk level in CORAS is represented as a value obtained from the *Likelihood* and *Consequence* of an *Unwanted Incident* and expressed in a scale such as *very low*, *low*, *high*

and *very high*. We will adopt this approach for the analysis of regrettable self-disclosure experiences and consequently for the derivation of privacy heuristics.

Surveillance Information The risks of self-disclosure are often grounded in the audience to which the information is being disclosed and the type of information being disclosed. Therefore, defining which information should be considered for privacy analysis is a very important aspect for the derivation of privacy heuristics. In the context of SNSs, privacy concerns related to data aggregation, probabilistic re-identification of individuals, as well as undesirable social categorizations ought to be discussed by the stakeholders [13]. This means that information that might not be personal per-se (e.g. potentially linkable data) can raise privacy concerns. Consequently, any observable information, regardless if that information can be linked to individuals, groups or communities, should be considered for privacy analysis. Such information, which covers Personally Identifiable Information (PII) and more, is defined by Gürses [13] as "surveillance information" (SI). Because of its broad scope, we will adopt this terminology for the identification and analysis of the information disclosed by the users of SNSs.

#	**Dimension**	**Surveillance Attributes**
I	Demographics	Age, Gender, Nationality, Racial origin, Ethnicity, Literacy level, Employment status, Income level, Family status
II	Sexual Profile	Sexual preference
III	Political Attitudes	Supported party, Political ideology
IV	Religious Beliefs	Supported religion
V	Health Factors and Condition	Smoking, Alcohol drinking, Drug use, Chronic diseases, Disabilities, Other health factors
VI	Location	Home location, Work location, Favorite places, Visited places
VII	Administrative	Personal Identification Number
VIII	Contact	Email address, Phone number
IX	Sentiment	Negative, Neutral, Positive

Table 1. The "self-disclosure" dimensions.

Self-disclosure Dimensions Equally important as the SI disclosed by the users, are the attributes enclosed in it. Petkos et al. [21] propose a taxonomy of personal data based on legal notions of personal information, as well as general perceptions of privacy and other state of the art definitions. This approach consists in organizing the user's private or sensitive personal attributes into different high-level categories called "privacy dimensions" (i.e. demographics, psychological traits, sexual profile, political attitudes, religious beliefs, health

factors and condition, location, and consumer profile). This taxonomy, unlike other approaches that focus on the source of the data (e.g. Schneider et al. [24]), has a strong focus on the semantics of the data about the user and allows a semantic and intuitive representation of different aspects of the user's personal information [21]. Many of these dimensions keep a strong correlation with the regrettable scenarios reported by the users in the study conducted by Wang et al. [30] discussed in Section 3.2 (e.g. users reported that sharing information about their religious beliefs and profanity had lead them to a regrettable experience). Consequently, based on the regret categories proposed by Wang et al. and taking into account the concept of SI, we have refined the original privacy dimensions of Petkos et. al. into what we call the "self-disclosure dimensions". These self-disclosure dimensions (Table 1), which are expressed as a set of "surveillance attributes" (SAs), allow us to analyze from a regret-oriented perspective the SI disclosed by the user in a post. Since the original categories were not covering attributes like email address, phone number, personal identification number [2] and sentiment, we added three new dimensions (namely Administrative, Contact and Sentiment) to the original taxonomy.

4.2 Method

The PHeDer method consists of four sequential steps which are *regret acknowledgment*, *concern analysis*, *heuristics design*, and *constraint integration*. As depicted in Fig. 3, each stage of the method draws on different external inputs and generates the outputs for the next step. The final output of the method is an updated version of the IAS'PHDB.

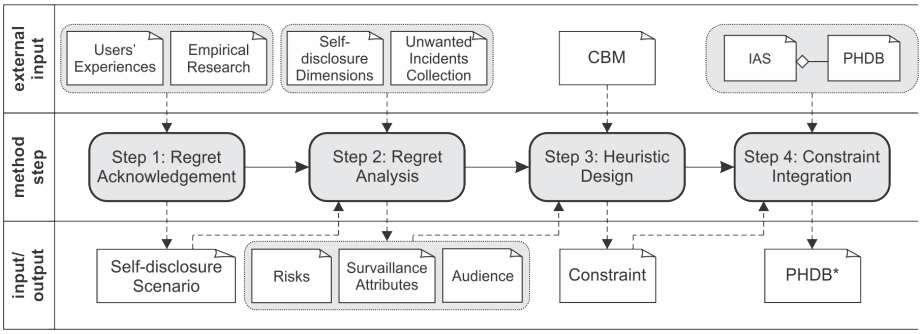


Fig. 3. PHeDer Steps and Artifacts

Step 1: Regret Acknowledgment The input for this step could be any evidence source of regret. Such evidence might come from regrettable experiences that the users reported themselves, or as the outcome of an empirical research like

[2] Examples of personal identification number are Social Security Number (SSN), passport number, drivers license number, taxpayer identification number, or financial account or credit card number [19].

the one conducted by Wang et al. [30]. For the sake of simplicity, we assume that a single development group carries forward all the steps of the method and counts with the results of an empirical study about regrettable experiences. However, since these experiences can take place in any moment in time, it would be convenient to provide "offline" communication channels (i.e. outside of an empirical research instance) to the users for direct communication with the development team. In this step, a regrettable scenario should be described informally by the development team in terms of which information was disclosed, which was the unintended audience that it reached, and what where the unwanted incidents that lead the user to a feeling of regret. The output of this step can be represented as in Fig. 4 which describes a scenario where a user reported that he/she regretted to write a negative comment about his/her workplace in a public post.

USER'S POST

"A typical day at the office. Lots of complaints and bad mood. Cannot wait for the day to be over...!"

Actual Audience: PUBLIC.
Unintended Audience: The user's work colleagues, or superior.
Unwanted Incidents: Wake-up call from superior; bad image; job loss.

Fig. 4. Example of self-disclosure scenario

Step 2: Regret Analysis The post shared by the user in the example of Fig. 4 contains information related to his/her *employment status* and *work location*, together with a *negative* sentiment. According to Table 1, these are SAs of the *demographics*, *location* and *sentiment* self-disclosure dimensions respectively. Therefore, it is possible to trace a semantic correlation between the content of the post and one or more self-disclosure dimensions, and consequently express a regrettable scenario in terms of one or more SAs.

As previously mentioned, a regrettable scenario can lead to one or more unwanted incidents with a certain likelihood of occurrence (i.e. a risk). Consequently, a risk function must be defined to estimate the likelihood and the impact of the unwanted incidents of a regrettable scenario. Like in CORAS, such a function can be represented as a matrix similar to the one in Table 2. This matrix is divided in four sections, each representing one of the risk levels: *very low* (green), *low* (yellow), *high* (orange) and *very high* (red). A risk level is derived from the frequency of the unwanted incident (i.e. rare, unlikely, possible, likely or certain) and its consequence (i.e. insignificant, minor, moderate, major or catastrophic). We assume that knowledge about unwanted incidents which can or have occurred as consequence of online self-disclosure are stored in an "Unwanted Incidents Collection". Such a collection will help to build the risk matrix and consequently to analyze the potential risks of a regrettable scenario.

Let us assume that the scenario described in Fig. 4 by the development team has three unwanted incidents *wake up call from superior (I1)*, *bad image (I2)*, and

		Consequence				
		Insignificant	*Minor*	*Moderate*	*Major*	*Catastrophic*
Likelihood	*Rare*					
	Unlikely					*I3*
	Possible				*I1*	
	Likely			*I2*		
	Certain					

Table 2. Example of risk matrix.

job loss (I3). One can consider that the frequency of such incidents is the same for every user in a SNS, and can therefore be determined in a global scale by a risk expert. Nevertheless, when it comes to the estimation of the consequences of each incident, global assumptions are harder to make. This is basically because, as mentioned in Section 4.1, users do not perceive the consequences of a self-disclosure act in the same levels. For instance, a bad image incident can be catastrophic for a certain user or group of users, or can be insignificant for others. Therefore, a risk matrix must be elaborated for every regrettable scenario and for every user or group of users with similar characteristics.

Clearly, to keep an individual risk matrix for every user is an unpractical and not efficient solution. Besides, different users can share the same severity perceptions towards a particular risk, meaning that they share the same privacy attitudes. Such similarities have been acknowledged by Westin who developed a "Privacy Segmentation Index" to categorize individuals into three privacy groups: *fundamentalists*, *pragmatists*, and *unconcerned* [32]. Privacy *fundamentalists* are at the maximum extreme of privacy concerns being the most protective of their privacy. Privacy *pragmatists* on the other hand evaluate the potential pros and cons of sharing information and make their decisions according to the trust they perceive towards the information's receiver. On the other extreme, privacy *unconcerned* are the less protective of their privacy since they perceive that the benefits of information disclosure far outweigh the potential negative consequences. These categories have been widely used to measure privacy attitudes and therefore could be beneficial for the elaboration of the risk matrix of regrettable scenarios. Users could be grouped into these three categories, which means that it would only be necessary to elaborate three risk matrices (one for each privacy attitude).

Step 3: Heuristic Design This step consists in the codification of the outcome of Step 2 (risk matrix, SAs, and audience) into privacy heuristics. According to Díaz Ferreyra et al. [8], the domain knowledge of IAS should be encoded following principles of Constraint Based Modeling (CBM) which postulates that domain

knowledge (i.e. privacy heuristics) can be represented as constraints on correct solutions of a problem (i.e. a self-disclosure scenario). Such correct solutions must satisfy a set of fundamental domain principles (encoded in constraints) that should not be violated. As long as the users never reach a state that is known to be wrong (i.e. a regrettable scenario), they are free to perform whatever actions they please. In this sense, a state constraint is a pair of *relevance* and *satisfaction* tests on a problem state, where each member of the pair can be seen as a set of features or properties that a problem state must satisfy [8].

In Snippet 1, *relevance* and *satisfaction* tests over a problem state are expressed as Horn Clauses in Prolog. The relevance condition consists of the left hand side of the *share* predicate, which acknowledges and evaluates an information disclosure event (in this case a post). Such event is modeled by the parameters [X|Xs] (a list of SAs where X is the first element), Au (the post's audience), and Usr (the user's id). Likewise, the satisfaction condition (right hand side of the predicate) evaluates the existence of a potential regrettable scenario associated with the disclosure of such SAs to a certain audience. In order to find out if the user's disclosure can derive in a regrettable scenario, the potential risks of the disclosure must be evaluated. This evaluation is carried out by the *regret* predicate which checks if there is an unwanted incident whose risk is not acceptable for the user. Since the risk acceptance depends on the user's privacy attitude, it is necessary to instantiate the Att variable with one of the *fundamentalist*, *pragmatist* or *unconcerned* values. This unification process consists of binding the content of the Att variable with an *attitude* predicate containing the same user's id. Following the same unification approach, the *srv_att_list* checks if [X|Xs] is not an empty list, and if it is composed by SAs.

```
share([X|Xs], Au, Usr):- srv_att_list([X|Xs]), audience(Au), user(Usr),
  attitude(Usr, Att), not regret([X|Xs], Au, Att).

regret([X|Xs], Au, Att):- unwanted_inc([X|Xs], Au, Att, Unwi),
  risk(Att, Unwi, Type, Cons, Freq, Level), not acceptable(Att, Level).

unwanted_inc([X|Xs], Au, Att, Unwi):- unw_incident([Y|Ys], Au, Att, Unwi),
  subset([Y|Ys],[X|Xs]).

srv_att_list([X]):- srv_att(X).
srv_att_list([X|Xs]):- srv_att(X), srv_att_list(Xs).
```

Snippet 1. Relevance and satisfaction conditions

Depending on the user's attitude, the impact of an unwanted incident can vary between *insignificant* and *catastrophic*. Therefore, the acceptance level of an unwanted incident also fluctuates between very low, low, high and very high, depending on the user's attitude. The *regret* predicate models the evaluation of the risks associated with the user's disclosure (i.e. the post) by taking into account his/her privacy attitude (Att), the list of SAs ([X|Xs]) and the audience (AU). First, the predicate invokes the *unw_incident* predicate, in order to find an

unwanted incident (i.e. instantiate the Unwi variable) linked with the SAs disclosed in the user's post, his/her attitude, and the post's audience. Thereafter, the *risk* predicate is invoked with the attitude and unwanted incident as parameters (Att and Unwi respectively) to compute the risk level of the unwanted incident (i.e. unify the Level variable). If the risk level of an unwanted incident is not acceptable according to the user's attitude, then the post is considered as potentially regrettable. Therefore, the last step of the *risk* predicate consists on checking the risk's acceptance. This is done by matching the unified variables Att and Level with an *acceptable* fact which defines the acceptance level of risk for each privacy attitude. For this, we assume that for a fundamentalist only very low risks are acceptable, for a pragmatist very low and low risks, and for a unconcerned the risks which are very low, low and high. If the risk is not acceptable, then the user's disclosure is assessed as a potential *regret* and the satisfaction condition of the *share* predicate gets violated.

```
unw_incident([Employmentstatus, Worklocation, Negative], Work, Job_loss).
risk(Pragmatist, Job_loss, Relational, Catastrophic, Rare, High).

audience(Work).
user(John).
attitude(John, Pragmatist).
acceptable(Pragmatist, Low).
acceptable(Pragmatist, Very_low).
srv_att(Worklocation).
srv_att(Negative).
srv_att(Employmentstatus).
```

Snippet 2. Privacy heuristic example

In order to asses our disclosure scenario, a set of facts which encode one or more privacy heuristics are evaluated. The heuristic of Snippet 2 has been derived from the analysis performed over the regrettable scenario described in Fig. 4. Here, the content of the risk matrix is encoded in the facts *unw_incident* and *risk*. The first one states that a job loss is an unwanted incident which occurs if SAs related to the user's employment status and work location together with a negative sentiment are disclosed to an audience containing people from his/her workplace. The second one states that such unwanted incident (that can be cataloged as Relational according to the categories described in 4.1) is rare to occur, but has a catastrophic impact among users with a pragmatic privacy attitude. Consequently, the risk is assessed as "high" for pragmatic users. Therefore, if a user John, who is a pragmatist, shares "A typical day at the office. Lots of complaints and bad mood. Cannot wait for the day to be over...!", then the risk is evaluated as not acceptable and the post considered as potentially regrettable.

Step 4: Constraint Integration Once the constraints are derived, we proceed to their incorporation in a PHDB like the one in IAS. As it is shown in

the Fig. 3, the association between PHDB and IAS is "weak", meaning that the PHDB does not compleately depend on an IAS. This is because a PHDB can serve other purposes which are not necessarily the ones of IAS (e.g. other awareness or privacy recommender systems with similar characteristics). On the other hand, it will depend on the particular implementation of the data base on how the integration procedure is executed. If the PHDB is encoded in Prolog as in the example, then the command *asserta* can be used to incorporate new facts and predicates to the data base[10]. Nevertheless, different implementations will require specific solutions for this step.

5 Privacy Heuristics Evaluation in IAS

Once an iteration of the PHeDer method is completed, a new set of privacy heuristics are included in the PHDB of an IAS. As described in Section 3.3, an IAS uses the knowledge stored in the PHDB and the UPDB in order to deliver a feedback message to the user when he/she is about to disclose a piece of SI in a post. The Algorithm 1 (function *AnalyzePost*) describes how this process is executed at run time. First, a *DetectSurvAtt* function (line 2) is in charge of tracing a semantic correlation between the content of the post and one or more SAs. This can be achieved for example by using Support Vector Machines for developing a classifier which automatically derives the SAs contained in a post (similar to the proposal of Nguyen-Son et. al. [20]). Once the post is expressed as a set of SAs, a *Share* function (like the one described in Snippet 1) assesses the potential risks of the disclosure and evaluates the scenario as *regrettable* or not (see line 5). If the post is considered as potentially regrettable for the user, then a feedback message must be raised informing about the risks of the disclosure and a set of possible actions to overcome this issue (for instance, hints on how to constraint the post's audience).

As explained in the previous section, both risk level and the level of acceptance depend on the user's privacy attitude. Therefore, the user's attitude is retrieved by the *GetUsrAttitude* function (line 7) to be later used by the *GetUnacRisks* to compute the set of unacceptable risks (line 8). For this, *GetUnacRisks* takes into account the SAs contained in the post, and the targeted audience in addition to the user's privacy attitude. Both functions, *GetUnacRisks* and *GetUsrAttitude*, can be easily implemented by querying the content of the PHDB. This is, using the predicates and facts of Snippet 1 and 2. Since the feedback must take into account how the user is performing regarding his/her privacy attitudes, a *GetUsrPerformance* function (line 9) collects such information from the UPDB as described in Section 3.3. The feedback generation concludes after calling the *GenFeedback* function (line 10), which taking into account the user's attitude, performance and unacceptable risks elaborates a tailored feedback message to the user. An implementation assessment for the generation of adaptive feedback goes beyond the scope of this paper and will be part of future work.

The study of Schäwel and Krämer [23] suggests that users of SNSs would engage with a system which holds the adaptive properties of IAS. Therefore,

Algorithm 1 Pseudo-code of the AnalyzePost algorithm

```
1: function AnalyzePost(Post P, Audience Au, User Usr)
2:     Set[SurvAttr] SAs := DetectSurvAtt(P);
3:     String feedbackMsg;
4:     if SAs ≠ ∅ then
5:         bool regrettable := ¬Share(SAs, Au, U);
6:         if regrettable then
7:             Attitude Att := GetUsrAttitude(U);
8:             Set[Risk] Rsks := GetUnacRisks(SAs, Au, Att);
9:             Performance Perf := GetUsrPerformance(U);
10:            feedbackMsg := GenFeedback(Perf, Rsks, Att);
11:        end if
12:    end if
13:    return feedbackMsg;
14: end function
```

an implementation of IAS needs to measure the effectiveness of the heuristics and consequently of the PHeDer method in the practice. Considering that self-disclosure is an activity which can take place across different SNSs, and many of them like Facebook offer an API for connecting to its services, an application for smartphones (app) is a good implementation option. Having a prototype of such app, a use case scenario with a group of users can be set up in order to evaluate their privacy attitudes before and after using an IAS. Consequently, in-depth interviews can be conducted to get more insights about the user's reactions and acceptance of the recommendations. This evaluation stage is part of an ongoing work in progress and is expected to be extensively discussed in a future research study.

6 Discussion and Future Work

One of the drawbacks of some adaptive preventative technologies like the one from Caliki et al. [6] is that privacy knowledge is learned from the user's previous disclosures (i.e. in a "supervised learning" approach). This means that new users of the system will spend some time without support until the first set of privacy rules is learned. This leads to another drawback which is that such approaches also rely in the assumption that the user's sharing decisions (i.e. training set) where always consistent with his/her privacy norms (i.e. the user has never accidentally revealed content to an unintended audience). Since this is not always the case, these systems are likely to learn wrong sharing rules in a non-controlled operational environment. To overcome this issue, the PHeDer method could be applied to generate a privacy knowledge base-line so that new users can have support from the very beginning, develop a proactive behavior, and consequently make fewer mistakes when sharing their information.

On the other hand, PHeDer relies in the assumption that users can be clustered according to their privacy attitudes like proposed by Westin. Current

research by Woodruff et al. has put the predictive potential of Westin's categories into question [33]. Basically, Westin's Privacy Segmentation Index consists of three questions and a set of rules to translate the answers into the three categories discussed in Section 4.2. However, these questions examine privacy attitudes about consumer control, business, laws, and regulations. Therefore, they capture broad generic privacy attitudes, which are not good predictors of context-specific privacy related behaviors. Moreover, the index seems to rely on the unstated assumption that individuals make privacy decisions that are highly rational, informed and reflective. This has been already questioned and documented in the so called "Privacy Paradox"[3] which revealed peoples' privacy attitude-behavior dichotomy. Consequently, and as suggested by Woodruff et. al., future work should consider alternative instruments to better capture and predict the users's privacy attitudes such as the Internet Users' Information Privacy Concern (IUIPC) scale [18] or the Privacy Concern Scale (PCS) [5].

Another possible critic to PHeDer is that the method is executed offline (not at run-time) and requires a study about users' regrettable disclosures as input. This hinders the incorporation of new heuristics into the PHDB, basically because of the cost of conducting such type of studies. This is, the time and the resources needed to recruit the participants of the study, as well as for data conditioning and the application of the method's steps. Thus, a run-time approach for learning this privacy heuristics would be beneficial for keeping up to date the content of the PHDB. One possible way is to examine the deleted posts of a user in terms of the disclosed SAs. If such post contains one or more SAs, then it could be considered as a regret. Of course, then the question arises about which were the reasons (unwanted incidents) that made the user delete the post. A simple solution would be to ask directly to the user this question and try to estimate the risks. Such a run-time approach for learning privacy heuristics is also part of our future work.

7 Conclusion

Since online self-disclosure takes place at run-time and not prior to the system's development phase, regrettable experiences are hardly taken into consideration for shaping privacy requirements. Consequently, the implementation of awareness mechanisms which satisfy such privacy requirements is often neglected. The method presented in this work considers users' regrettable experiences as explicit manifestations of privacy concerns. Therefore, it can be seen as a user-oriented elicitation method of privacy requirements for SNSs. Consequently, the heuristics derived from the method can not only shape better awareness mechanisms and preventative technologies like IAS, but also improve the ones in the state of the art. We believe that using heuristics derived from the users' regrets to raise awareness is promising not only for promoting a proactive privacy behavior, but also for making the tie between the user and his/her digital data more emotionally appreciable. It is matter of future research to evaluate the effectiveness of

such heuristics in a prototype of IAS, as to develop engagement mechanisms for making privacy awareness an ongoing and sustained learning process.

Acknowledgments. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

References

1. Ackerman, A.: Visceral Reactions: Emotional Pay Dirt or Fast Track to Melodrama? (May 2012), retrieved March 2, 2017 from http://www.helpingwritersbecomeauthors.com/visceral-reactions-emotional-pay-dirt/
2. Archer, R.L.: Self-disclosure. In: The self in social psychology, pp. 183–204. Oxford University Press (March 1980)
3. Barnes, S.B.: A privacy paradox: Social networking in the United States. First Monday 11(9) (September 2006), http://dx.doi.org/10.5210/fm.v11i9.1394
4. Boyd, D.M., Ellison, N.B.: Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication 13(1), 210–230 (October 2007), http://dx.doi.org/10.1111/j.1083-6101.2007.00393.x
5. Buchanan, T., Paine, C., Joinson, A.N., Reips, U.D.: Development of measures of online privacy concern and protection for use on the internet. Journal of the American Society for Information Science and Technology 58(2), 157–165 (November 2007), http://dx.doi.org/10.1002/asi.20459
6. Calikli, G., Law, M., Bandara, A.K., Russo, A., Dickens, L., Price, B.A., Stuart, A., Levine, M., Nuseibeh, B.: Privacy Dynamics: Learning Privacy Norms for Social Software. In: Proceedings of the 11th International Symposium on Software Engineering for Adaptive and Self-Managing Systems. pp. 47–56. ACM (May 2016)
7. Diaz, C., Gürses, S.: Understanding the landscape of privacy technologies (extended abstract). In: Proceedings of the Information Security Summit, ISS 2012. pp. 58–63 (May 2012)
8. Díaz Ferreyra, N.E., Schäwel, J., Heisel, M., Meske, C.: Addressing Self-disclosure in Social Media: An Instructional Awareness Approach. In: Proceedings of the 2nd ACS/IEEE International Workshop on Online Social Networks Technologies (OSNT). ACS/IEEE (December 2016)
9. Fang, L., LeFevre, K.: Privacy wizards for social networking sites. In: Proceedings of the 19th International Conference on World Wide Web. pp. 351–360. WWW '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1772690.1772727
10. Frühwirth, T., De Koninck, L., Triska, M., Wielemaker, J.: SWI Prolog Reference Manual 6.2. 2. BoD–Books on Demand (2012)
11. Ghazinour, K., Matwin, S., Sokolova, M.: YourPrivacyProtector: A Recommender System for Privacy Settings in Social Networks. International Journal of Security, Privacy and Trust Management (IJSPTM) 2(4) (August 2013)
12. Group, O.M.: OMG Unified Modeling Language (OMG UML). OMG Document Number formal/2015-03-01 (March 2015)
13. Gürses, S.: Multilateral Privacy Rquirements Analysis in Online Social Networks. Ph.D. thesis, KU Leuven, Heverlee (2010)

14. Gürses, S., Rizk, R., Gunther, O.: Privacy Design in Online Social Setworks: Learning from Privacy Breaches and Community Feedback. In: Proceedings of the International Conference on Information Systems, ICIS 2008. p. 90 (December 2008)
15. Kietzmann, J.H., Hermkens, K., McCarthy, I.P., Silvestre, B.S.: Social media? get serious! understanding the functional building blocks of social media. Business Horizons 54(3), 241–251 (May 2011), http://dx.doi.org/10.1016/j.bushor.2011.01.005
16. Krämer, N., Haferkamp, N.: Online self-presentation: Balancing privacy concerns and impression construction on social networking sites. In: Privacy Online, pp. 127–141. Springer (2011)
17. Lund, M.S., Solhaug, B., Stølen, K.: Model-Driven Risk Analysis: The CORAS Approach. Springer Science & Business Media (October 2010)
18. Malhotra, N.K., Kim, S.S., Agarwal, J.: Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. In: Information Systems Research, vol. 15, pp. 336–355. Informs (December 2004)
19. McCallister, E., Grance, T., Scarfone, K.A.: Guide to protecting the confidentiality of Personally Identifiable Information (PII). DIANE Publishing (2010)
20. Nguyen-Son, H.Q., Tran, M.T., Yoshiura, H., Sonehara, N., Echizen, I.: Anonymizing Personal Text Messages Posted in Online Social Networks and Detecting Disclosures of Personal Information. IEICE TRANSACTIONS on Information and Systems 98(1), 78–88 (January 2015)
21. Petkos, G., Papadopoulos, S., Kompatsiaris, Y.: PScore: A Framework for Enhancing Privacy Awareness in Online Social Networks. In: Proceedings of the 10th International Conference on Availability, Reliability and Security, ARES 2015. pp. 592–600. IEEE (August 2015)
22. Petronio, S.: Boundaries of Privacy: Dialectics of Disclosure. Suny Press (February 2012)
23. Schäwel, J., Krämer, N.: Paving the Way for Technical Privacy Support: A Qualitative Study on Users' Intentions to Engage in Privacy Protection. In: The 67th Annual Conference of the International Communication Association (2017)
24. Schneier, B.: A taxonomy of social networking data. IEEE Security and Privacy 8(4), 88–88 (July 2010)
25. Stark, L.: The Emotional Context of Information Privacy. The Information Society 32(1), 14–27 (January 2016)
26. Trepte, S., Teutsch, D., Masur, P.K., Eicher, C., Fischer, M., Hennhöfer, A., Lind, F.: Do People Know about Privacy and Data Protection Strategies? Towards the "Online Privacy Literacy Scale"(OPLIS). In: Reforming European Data Protection Law, pp. 333–365. Springer Netherlands (2015)
27. Utz, S., Krämer, N.: The privacy paradox on social network sites revisited: The role of individual characteristics and group norms. Cyberpsychology: Journal of Psychosocial Research on Cyberspace 3(2) (2009)
28. Vitak, J.: Balancing privacy concerns and impression management strategies on Facebook. In: Proceedings of the Eleventh Symposium on Usable Privacy and Security, SOUPS 2015. USENIX (July 2015)
29. Wang, Y., Leon, P.G., Scott, K., Chen, X., Acquisti, A., Cranor, L.F.: Privacy Nudges for Social Media: An Exploratory Facebook Study. In: Proceedings of the 22nd International Conference on World Wide Web. pp. 763–770. ACM (2013)
30. Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P.G., Cranor, L.F.: I regretted the minute I pressed share: A Qualitative Study of Regrets on Facebook. In: Proceedings of the Seventh Symposium on Usable Privacy and Security, SOUPS 2011. ACM (2011)

31. Wang, Y.C., Burke, M., Kraut, R.: Modeling self-disclosure in social networking sites. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW 2016. pp. 74–85. ACM (February 2016)
32. Westin, A.F.: Privacy and Freedom. Washington and Lee Law Review 25(1), 166 (January 1968)
33. Woodruff, A., Pihur, V., Consolvo, S., Schmidt, L., Brandimarte, L., Acquisti, A.: Would a privacy fundamentalist sell their dna for $1000... if nothing bad happened as a result? the westin categories, behavioral intentions, and consequences. In: Proceedings of the Tenth Symposium on Usable Privacy and Security, SOUPS 2014. USENIX (2014)

DECLARATION OF INDIVIDUAL CONTRIBUTIONS

Publication title: Towards an ILP Approach for Learning Privacy Heuristics from Users' Regrets

Reference item: N. E. Díaz Ferreyra, R. Meis, and M. Heisel, "Towards an ILP Approach for Learning Privacy Heuristics from Users' Regrets," in *Network Intelligence Meets User Centered Social Media Networks. ENIC 2017*. Springer LNSN, August 2018, pp. 187–197. doi: 10.1007/978-3-319-90312-5_13

Author	Contribution	%
Nicolás E. Díaz Ferreyra	- Conceptualisation of the approach. - Planification of the work. - Draft of the manuscript.	85%
Rene Meis	- Discussion of the approach.	10%
Maritta Heisel	- Supervision and advice.	5%

Nicolás E. Díaz Ferreyra

Rene Meis

Maritta Heisel

Towards an ILP Approach for Learning Privacy Heuristics From Users' Regrets

Nicolás Emilio Díaz Ferreyra, Rene Meis, and Maritta Heisel

University of Duisburg Essen, Germany
{nicolas.diaz-ferreyra,rene.meis,maritta.heisel}@uni-due.de
https://www.ucsm.info/

Abstract. Disclosing private information in Social Network Sites (SNSs) often derives in unwanted incidents for the users (such as bad image, identity theft or unjustified discrimination), along with a feeling of regret and repentance. Regrettable online self-disclosure experiences can be seen as sources of privacy heuristics (best practices) that can help shaping better privacy awareness mechanisms. Considering deleted posts as an explicit manifestation of users' regrets, we propose an Inductive Logic Programming (ILP) approach for learning privacy heuristics. In this paper we introduce the motivating scenario and the theoretical foundations of this approach, and we provide an initial assessment towards its implementation.

Keywords: adaptive privacy, self-disclosure, awareness, social network sites, inductive logic programming

1 Introduction

Privacy norms play an important role in the construction of people's identity on an individual and collective level [3]. However, users of Social Network Sites (SNSs) like Facebook or Twitter seem to deliberately violate their privacy norms while interacting within these media platforms [2]. This is, users often share private or sensitive information that reaches unintended audiences in many cases. Consequently, users sharing acts derive in unwanted incidents (like stalking, identity theft, sextortion and job loss), along with a feeling of regret [4].

As one can see, there are two central aspects of information sharing: (i) not every information is appropriate to be mentioned in a particular context, and (ii) the audience to which information is disclosed influences such context. For instance, a user can consider acceptable to share his/her political affiliation with his family, but inappropriate with his/her work colleagues. Such aspects, which have been already identified in privacy theories (e.g. "contextual integrity" by Nissembaum [10]), are basic for the definition of privacy norms and consequently for individual privacy control.

In order to create a more privacy aware social environment, media technologies should support the users by providing guidance and feedback on their disclosures. Moreover, awareness mechanisms should be able to identify and learn

from regrettable experiences in order to provide effective privacy support. In line with these premises, Diaz Ferreyra et al. introduced the concept of Instructional Awareness System (IAS) which uses a privacy heuristics data base in order to generate adaptive awareness [5][4]. In this paper we propose to take a closer look into users' deleted posts and consider them as potential sources of privacy heuristics. Using Inductive Logic Programming (ILP) as the learning approach and deleted posts as the training set, we provide an initial assessment towards a privacy heuristics inference engine for IAS.

The content of this paper is organized as follows. In the next section, we discuss related work. Section 3 introduces the basic notions of ILP that form the paper's background. Following, in Section 4, we discuss how regrets can be identified and retrieved from users' deleted posts. In Section 5, we introduce a representation for regrettable scenarios and the initial components of an ILP system for learning privacy heuristics. Finally, we conclude in Section 6 with an outline and directions for future work.

2 Related Work

In the context of SNSs, adaptive privacy awareness systems seek to provide tailored feedback to the users when they attempt to reveal private information inside a post or in their profiles. Such feedback normally takes the form of a warning message that is displayed when a potentially unwanted disclosure is detected (e.g. a user revealing his/her bank account details in a public post on Facebook). For instance, Caliki et al. [2] developed a system which learns privacy rules from the user's previous sharing history, to use them later on as a criterion for raising awareness. This is, under the assumption that the user has never revealed private content to an unintended audience, the system infers *allow/deny(data, audience)* privacy rules though a machine learning engine (where "data" is a piece of private information like *bank account*, and "audience" is a sub-group of the user's Facebook friends). If a rule derived during the learning stage of the system is later violated in its operational stage, then a warning message is displayed. In line with this approach, Fang and LeFevre [6] introduced a system which analyses the information disclosed inside the user's Facebook profile in order to derive privacy preferences. This system recommends which attributes of the profile should be visible or not within a certain group of friends.

Following similar adaptivity principles, Díaz Ferreyra et al. developed an architecture which prescribes the basic components of an Instructional Awareness System (IAS) [5]. Using privacy best practices stored inside a Privacy Heuristics Data Base (PHDB), IAS generates a personalized warning message when a user attempts to reveal private data in a SNS post. The privacy heuristics in the PHDB are the outcome of a method which analyses the risks of regrettable online experiences reported by the users. This is, the input of the method are the experiences that users have reported themselves (to the development team of IAS for instance), or the outcome of an empirical research (e.g. questionnaires

or face to face interviews). This approach is effective for building a baseline of heuristics prior to the execution of the system. However, eliciting new entries of the PHDB requires the execution of this process which can be expensive and inefficient in terms of the resources and time needed to conduct interviews and process the outcome of them. Therefore, a *run-time* (*online*) approach for learning privacy heuristics would be beneficial for keeping up to date the content of IAS's PHDB.

3 Theoretical Background

This work proposes to endow IAS with a Privacy Heuristics Learning Engine (PHLE) based in principles of ILP. In this section we introduce the theoretical foundations of ILP and discuss its potential for carrying forward this task.

3.1 The ILP Problem

ILP is a discipline which employs techniques from machine learning and logic programming to infer hypotheses H from a set of observations and some background knowledge B [8]. Such observations are a set of positive and negative examples $E = E^+ \cup E^-$ of a concept to learn, and like the background knowledge, they are expressed as logic programs. The goal of ILP is to generate a logic program where positive examples are satisfied and negative examples are not. Let us consider that E^+ is expressed as ground unit definite clauses and E^- as ground unit headless Horn clauses of a single target predicate [7][1]. Then E contains ground unit clauses of a single predicate, and we can specify the general ILP problem as $B \wedge H \models E$. Since the goal is to find the simplest hypothesis, each clause in H should explain at least one example. If we consider H and E as single Horn clauses, then this expression can be rearranged as $B \wedge \neg E \models \neg H$.

Let $\neg\bot$ be the (potentially infinite) conjunction of ground literals which are true in all models of $B \wedge \neg E$ [9]. Considering that $\neg H$ must be true in every model of $B \wedge \neg E$, then it must contain a subset of the ground literals in $\neg\bot$. This is, $B \wedge \neg E \models \neg\bot$ and $\neg\bot \models \neg H$. Rearranging this expression, we obtain $H \models \bot$, which means that a subset of the solutions for H can be derived from $\bot$ by generalizing each given example in E. This means, $\bot$ is a clause that θ-subsumes[2] each example $e \in E$ [9]. This way, the final $\bot$ can be constructed as the disjunction of the body literals of all these derived clauses [9]. However, since $\bot$ can have infinite cardinality, the search space of those clauses which imply $\bot$ can also be infinite. In order to bound the search space of consistent and complete hypotheses, $\bot$ is built and generalized using *mode declarations* [9][11].

[1] Let $e(X)$ be the predicate which defines the examples, and $L = L_1, ..., L_n$ a set of ground literals which subsume the variable X. Then, positive examples can be expressed as $e(L_i)$., and negative examples as : $-e(L_j)$, $\forall\ 1 \leq i, j \leq n$

[2] A clause c_1 θ-subsumes a clause c_2 if and only if there exists a substitution θ such that $c_1\theta \subseteq c_2$. Consequently c_1 is a generalization of c_2 (and c_2 specialization of c_1) under θ-subsumption [8]

3.2 Mode Declarations and Types

Mode declarations are *language bias* (i.e. syntactic restrictions) that are imposed on candidate hypotheses in order to make the search space more efficient [9]. A mode declaration has either the form $modeh(n, s)$ or $modeb(n, s)$ for the head or body of a rule, respectively[9]. Consequently, it restricts the predicates which can occur in the head and body of a rule. The schema s is a ground literal with placemarkers of the form '*+type*', '*-type*', and '*#type*' standing for input variables, output variables, and constants, respectively [9]. The value n is the *recall* of the mode declarations, and is used to bound the number of times the scheme can be used. Let us consider the following example:

```
:- modeh(1, fly(+bird))?
:- modeb(1, wings(+bird,#property,-int))?
property(has_flight_feathers).
```

The first mode declaration says that general rules may have heads containing the predicate $fly(X)$ where X is a type of *bird* [1]. The second says that general rules may have bodies containing the predicate $wings(X, has\_flight\_feathers, Y)$, where X is a type of *bird*, Y an *integer*, and $has\_flight\_feathers$ a *property* [1]. The value of the recall here is one; therefore both mode declarations can be used once in the construction of the bottom clause.

4 Regrets Identification in SNSs

As discussed in the previous section, ILP is a supervised machine learning approach that can serve the purpose of developing a Privacy Heuristics Learning Engine (PHLE) for IAS. One sub-task in the development of such PHLE is to generate a set of examples (training set) to provide as input to the hypotheses derivation process. Since in many cases the information enclosed within a post has a private or sensitive semantics, posts become elemental units for user-centered privacy analysis in SNSs. In this sense, performing a privacy analysis of a post consists of determining the existence of private information enclosed in it. Such privacy analysis of user-generated content requires a taxonomy of attributes that can be considered as private. For this purpose, Díaz Ferreyra et al. proposed to organize the users' private or sensitive personal attributes around different high-level categories called the "self-disclosure dimensions" (i.e. demographics, sexual profile, political attitudes, religious beliefs, health factors and condition, location, administrative, contact, and sentiment) [4]. Each dimension consists of a set of Surveillance Attributes (SAs)[3] which allow to analyze from a user-centered perspective the information disclosed inside a post. For instance, in the scenario described in Fig. 1, the SAs *employment status*, *work location*, and *negative* sentiment are disclosed inside a post.

[3] SAs are those which can be linked to an individual, groups or communities and can raise privacy concerns related to data aggregation, probabilistic re-identification and undesirable social categorizations [4].

USER'S POST *"What a lame environment at the office...I can only hear people complain! Thanks God it's Friday! #tgif"*

Actual Audience: PUBLIC.
Unintended Audience: The user's work colleagues, or superior.
Unwanted Incidents: Wake-up call from superior; bad image; job loss.

Fig. 1. Example of self-disclosure scenario

When one or more SAs reach an unintended audience, an unwanted incident can take place and derive in a feeling of regret. In Fig. 1, the post reached the user's work acquaintances causing a wake-up call from superior, together with a negative impact on the user's image, and eventual job loss. Since the likelihood and impact of such incidents define the risks of the given scenario, they can be used to model a regret. This is, regrets can be expressed in terms of the SAs, unintended audience, and the associated risks (i.e. likelihood and impact of an unwanted incident). Risks, likelihood and consequence, can be expressed in a nominal scale. In this sense, a consequence is a value on an impact scale such as insignificant, minor, moderate, major or catastrophic. Likewise, the likelihood is a value on a frequency scale such as rare, unlikely, possible, likely and certain. Finally, a risk is a value obtained from the likelihood and consequence of the unwanted incident and expressed in a scale such as very low, low, high and very high.

5 Towards the development of a Privacy Heuristics Learning Engine

So far we have discussed the concepts that serve in the identification of regrettable posts and therefore for creating examples for a PHLE. However, in practice, when a post is deleted there is not much information about the risks and, moreover, if the deleted post has derived or not in a regret for the user. In this section we provide an approach on how this missing information can be retrieved and used later on for the development of privacy heuristics. We also introduce a syntax for such heuristics in the IAS's PHDB together with the respective mode declarations for their automatic inference.

5.1 Privacy Heuristics

Snippet 1 describes a syntax for the entries in the PHDB, namely the hypotheses (i.e. regret conditions) to be learned by the PHLE. These learned hypotheses are the privacy heuristics which will help to identify potential regrettable scenarios in the future. That is, when a user attempts to disclose SAs inside a post, IAS should query the PHDB in order to verify that this will not lead to a potential regret. In other words, we want to find the consequence (Cons) and

frequency (Freq) of a potential unwanted incident (Unwi) based in the disclosed SAs ([X|Xs]) and the list of users that conform the post's audience ([Y|Ys]). The *regret* predicate models this conceptual relation, and is the entry point for querying the PHDB. Consequently, [X|Xs] and [Y|Ys] are input variables of a *regret-?* query, and Unwi, Cons and Freq are the output variables.

The impact of an unwanted incident (and consequently the acceptance level of the risk associated with it) are perceived differently among individuals. For instance, some users might consider the consequence associated with a "wake up call from supervisor" as Insignificant, and others as Catastrophic. This is because individuals have different privacy attitudes which influence the perception level of risky events [4]. In order to model this, we will adopt the Westin's Privacy Index for the classification of the users' privacy attitudes [12]. Such index classify individuals into three privacy groups: fundamentalists, pragmatists, and unconcerned (each group with high, medium and low levels of privacy concerns respectively). The predicate *acceptable*, represents this relation between the user's privacy attitude and the risk acceptance level associated with it.

```
/*Privacy Heuristic*/
regret([X|Xs], [Y|Ys], Unwi, Cons, Freq):-
  srv_att_list([X|Xs]), usr_list([Y|Ys]), inSIG([Y|Ys], SIG1),
  unwanted_inc([emp_status, work_location, neg], SIG1, wake_up_call),
  subset([X|Xs], [emp_status, work_location, neg]),
  risk(wake_up_call, Cons, Freq, Level), not acceptable(Att, Level).
```

Snippet 1. Example of a Privacy Heuristic

Another element of a regret is the audience towards which a set of SAs has been disclosed. Normally in SNSs like Facebook, a user has a list of "friends" composed by other users of Facebook. We will adopt this approach and assume that the user's friends list is grouped into different circles which are constructed under the premises of Social Identity Theory (SIT). Basically SIT postulates that people belong to multiple social identities (for instance, being Swedish, being an athlete, or being a researcher are all examples of identities/identity groups). Since users frequently have a mental conceptualization of the different social identity groups with whom they interact, we will assume that the friends list of a user is clustered into a set of identity groups as suggested by SIT. We can imagine for instance groups like *work*, *church*, *gym*, or *choir* depending on the social circles that a user belongs to. The predicate *inSIG*, evaluates if the audience to which the post has been disclosed corresponds to one of the social identity groups (SIG) in which the user's friends list is clustered.

To learn the circumstances of a regrettable scenario means to define instances of some of the variables which appear in the body of the *regret* predicate. Particularly, we are interested in knowing which concrete SAs lead to a certain unwanted incident when disclosed to a particular SIG. Consequently, the SAs, unwanted incident and SIG involved in the scenario of Fig. 1, are represented by the ground literals *[emp_status, work_location, neg]*, *wake_up_call*, and *SIG1* re-

spectively (we will consider *SIG1* as the identifier of the SIG "work place"). The relation between these elements is modeled by the *unwanted_inc* predicate, and the *subset* predicate verifies if the SAs involved in the unwanted incident are a subset of the SAs disclosed inside the deleted post. Likewise, the *risk* predicate models the semantic relation between the unwanted incident (*wake_up_call* in this case), its frequency and its consequence (as described in Section 4). As already mentioned, the risk of an unwanted incident (which is expressed as a level on a risk scale) can be acceptable or nor depending on the user's privacy attitude. For instance, we can assume that for fundamentalists only very low risks are acceptable, for a pragmatist very low and low risks, and for an unconcerned the risks which are very low, low and high. The acceptance of the risk level, will at the end determinate if the disclosure scenario which is being evaluated will lead to a potential state of regret or not.

5.2 Regrets Retrieval

Once the relevant SAs are identified, the information about the causes which motivated the user to delete the post have to be retrieved. Since this information is not evident at a glance, we propose to build an interface which asks questions to the user when a "delete" event takes place. That is, when a post is deleted, we first analyze if it contains SAs, and if so we ask the user for extra information to complete the description of the regrettable scenario. A mock-up of the described interface is depicted in Fig. 2, where the reported risk information corresponds to the post described in Fig. 1.

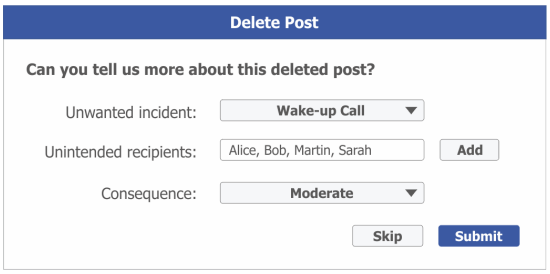


Fig. 2. "Delete Post" Interface

Among the information requested from the user there is the "unintended recipients" of the post. This corresponds to a list of users which were part of the original audience of the post, but should have not been included for privacy reasons. In the example, the user reports that the post reached his/her work colleagues Alice, Bob, Martin and Sarah, and has selected them as the unintended recipients. Likewise, the user reports that the unwanted incident has been a *wake-up call* and that the consequence has been perceived by him/her as *moderate*. With this information submitted by the user (i.e. unintended recipients, unwanted incident, and consequence), and the list of SAs extracted

from the post, we can create a *regret* predicate consisting of the ground literals: *[emp_status, work_location, neg]*, *[Alice, Bob, Martin, Sarah]*, *wake_up_call*, *moderate*, *certain*. This *regret* predicate represents a concrete regrettable scenario, and is therefore a training example for the PHLE.

Once the regret scenario is submitted by the user, IAS will first store the information about the risks. That is, it will generate a *risk* entry in the PHDB with the information about the unwanted incident and its consequence. Since a *risk* predicate contains also the frequency of the unwanted incident and the risk level, such values must be also generated. For the frequency, we will adopt "certain" for every case since we can assume that if a post containing SAs has been deleted is because a regret took place. Likewise, risks levels can be defined for every value combination of likelihood and consequence. We will assume that when the likelihood is certain and the consequence is moderate, the risk level is very high. Consequently, for our example, a *risk* predicate with the ground literals *wake_up_call*, *moderate*, *certain*, and *very_high* is created as a new entry in the PHDB. Risks together with the information of their acceptance level (the *acceptable* predicates), are part of the background knowledge of the PHDB. In this case we will consider that the user is a *pragmatist*, and therefore only risks which are *low* or *very_low* are acceptable (see Snippet 2).

```
/*Regret Example*/
regret([emp_status, work_location, neg], [Alice, Bob, Martin, Sarah],
    wake_up_call, moderate, certain).
/*Background Knowledge*/
risk(wake_up_call, moderate, certain, very_high).
sig([Bill, Bob, Sam, Sarah, Alice, John, Martin], SIG1).
acceptable(pragmatist, low).
acceptable(pragmatist, very_low).
```

Snippet 2. A regret example submitted by the user

Another element which is part of the background knowledge are the different SIGs in which the user's friend list is grouped. As can be observed in Snippet 2, the predicate *sig* assigns an identifier to each of the SIGs. For instance, SIG1 refers to a group of users consisting of *[Bill, Bob, Sam, Sarah, Alice, John, Martin]*. This information allows to derive heuristics where the audience is generalized to a SIG. This is, based on the examples of regrettable scenarios, identify the SIG to which a set of SAs should not be revealed to. In Snippet 2, we can see that the unintended recipients could be generalized to SIG1. Consequently, the next time the user attempts to disclose the SAs *[emp_status, work_location, neg]* to an audience containing a member of SIG1, then IAS will raise a warning message.

5.3 Mode Declarations and Type Definitions

In order to define the mode declarations, first it is necessary to define the *types* of the elements which are part of them. That, is to describe the categories of

objects (number, lists, names, etc) in the domain being modeled. In our case we need the types *Cons*, *Freq*, *Unwi*, *Usr*, *SA*, *Att*, *Level*, *srv_att_list*, *usr_list*. Due to space limitations we only provide a partial description of these in Snippet 3. The missing types can be defined in an analogous way. Once these types are defined, we can then proceed with the definition of the head and body mode declarations. The head of a heuristic (i.e. a *regret* predicate) is a function of objects of the types *srv_att_list*, *usr_list*, *Unwi*, *Cons*, and *Freq*. All of these objects are variables in the head of the *regret* predicate, however (as mentioned in Section 5.1), only *srv_att_list* and *usr_list* are input variables. Therefore, the placemarkers of the *modeh* for the head of the *regret* predicate will be *+srv_att_list*, *+usr_list*, *-Unwi*, *-Cons*, and *-Freq*.

```
/*Types*/
SA(emp_status). SA(work_location). SA(neg).
...
srv_att_list([]).
srv_att_list([X|Xs]):-SA(X), srv_att_list(Xs).
/*Mode declarations*/
:- modeh(1, regret(+srv_att_list, +usr_list, -Unwi, -Cons, -Freq))?
:- modeb(1, inSIG,(+usr_list, #SIG)?
:- modeb(1, unwanted_inc(#srv_att_list, #SIG, #Unwi))?
:- modeb(1, subset(+srv_att_list, #srv_att_list)?
:- modeb(1, risk(#Unwi, -Cons, -Freq, -Level))?
:- modeb(1, not acceptable(+Att, +Level))?
```

Snippet 3. Types and mode declarations

The body of the *regret* predicate is defined by the predicates *inSIG*, *unwanted_inc*, *subset*, *risk*, and *not acceptable*. Therefore, we need to define a body mode declaration *modeb* for each one of these predicates. Following the same criterion used to define the input and output variables of *modeh*, we define the arguments for each *modeb*. Since many of the predicates used in the body of *regret* are facts or contain grounded literals in their declaration, we express their respective *modeb* using the *#type* placemarker. Such is the case of the *risk* clause which in the body of the *regret* rule has its parameter *Unwi* defined as a constant. Likewise, the argument *SIG* in *inSIG*, *srv_att_list* in *subset*, and all the arguments of *unwanted_inc* are grounded literals in the body of *regret*. Therefore, we express them as constants in their respective body mode declarations.

6 Conclusion and Future Work

So far, we have introduced an ILP model of a PHLE which learns privacy heuristics from regrettable posts in SNSs. We believe that this is a promising approach which will contribute in shaping better and more user-centered preventative technologies. It is a matter of future research to develop a prototype of the model here introduced in order to measure its performance, as to evaluate if adaptive

awareness approaches like IAS indeed help to achieve better engagement levels. Other directions for future research involve the development of alternative approaches to Westin's Privacy Index for measuring the users' perceived severity of unwanted incidents (as suggested by Díaz Ferreyra et al. [4]). This involves the development of a user interface (in the form of a short questionnaire) to capture the users' attitudes towards privacy risks in order to improve the performance of the PHDB and consequently of IAS.

Acknowledgments. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

References

1. Athakravi, D., Broda, K., Russo, A.: Predicate invention in inductive logic programming. In: OASIcs-OpenAccess Series in Informatics. vol. 28. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2012)
2. Calikli, G., Law, M., Bandara, A.K., Russo, A., Dickens, L., Price, B.A., Stuart, A., Levine, M., Nuseibeh, B.: Privacy Dynamics: Learning Privacy Norms for Social Software. In: Proceedings of the 11th International Symposium on Software Engineering for Adaptive and Self-Managing Systems. pp. 47–56. ACM (May 2016)
3. Diaz, C., Gürses, S.: Understanding the landscape of privacy technologies (extended abstract). In: Proceedings of the Information Security Summit, ISS 2012. pp. 58–63 (May 2012)
4. Díaz Ferreyra, N.E., Meis, R., Heisel, M.: Online Self-disclosure: From Users' Regrets to Instructional Awareness. In: Proceedings of the IFIP International Cross-Domain Conference (CD-MAKE) (August 2017), Accapted for publication
5. Díaz Ferreyra, N.E., Schäwel, J., Heisel, M., Meske, C.: Addressing Self-disclosure in Social Media: An Instructional Awareness Approach. In: Proceedings of the 2nd ACS/IEEE International Workshop on Online Social Networks Technologies (OSNT). ACS/IEEE (December 2016)
6. Fang, L., LeFevre, K.: Privacy wizards for social networking sites. In: Proceedings of the 19th International Conference on World Wide Web. pp. 351–360. WWW '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1772690.1772727
7. Muggleton, S.: Inverse Entailment and Progol. New Generation Computing 13(3), 245–286 (1995)
8. Muggleton, S., De Raedt, L.: Inductive Logic Programming: Theory and Methods. The Journal of Logic Programming 19, 629–679 (1994)
9. Muggleton, S.H., Firth, J.: CProgol4.4: a tutorial introduction. In: Dzeroski, S., Lavrac, N. (eds.) Relational Data Mining, pp. 160–188. Springer-Verlag (2001), http://www.doc.ic.ac.uk/\~shm/Papers/progtuttheo.pdf
10. Nissenbaum, H.: Privacy as contextual integrity. Wash. L. Rev. 79, 119 (2004)
11. Roberts, S.: An introduction to progol. Department of Computer Science, University of York 244 (1997)
12. Westin, A.F.: Privacy and Freedom. Washington and Lee Law Review 25(1), 166 (January 1968)

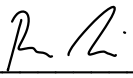
DECLARATION OF INDIVIDUAL CONTRIBUTIONS

Publication title: Should User-generated Content be a Matter of Privacy Awareness? A position paper

Reference item: N. E. Díaz Ferreyra, R. Meis, and M. Heisel, “Should User-generated Content be a Matter of Privacy Awareness? A position paper,” in *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 3: KMIS.* SCITEPRESS, November 2018, pp. 212–216. doi: 10.5220/0006517302120216

Author	Contribution	%
Nicolás E. Díaz Ferreyra	- Conceptualisation of the approach. - Planification of the work. - Draft of the manuscript.	90%
Rene Meis	- Supervision and advice.	5%
Maritta Heisel	- Supervision and advice.	5%

Nicolás E. Díaz Ferreyra



Rene Meis

Maritta Heisel

Should User-generated Content be a Matter of Privacy Awareness?
A position paper

Nicolás Emilio Díaz Ferreyra, Rene Meis and Maritta Heisel
University of Duisburg-Essen, Germany
RTG User-Centred Social Media
https://www.ucsm.info/
{*nicolas.diaz-ferreyra, rene.meis, maritta.heisel*}*@uni-due.de*

Keywords: adaptive privacy, self-disclosure, awareness, social network sites, data visceralization

Abstract: Social Network Sites (SNSs) like Facebook or Twitter have radically redefined the mechanisms for social interaction. One of the main aspects of these platforms are their information sharing features which allow user-generated content to reach wide and diverse audiences within a few seconds. Whereas the spectrum of shared content is large and varied, it can nevertheless include private and sensitive information. Such content of sensitive nature can derive in unwanted incidents for the users (such as reputation damage, job loss, or harassment) when reaching unintended audiences. In this paper, we analyse and discuss the privacy risks of information disclosure in SNSs from a user-centred perspective. We argue that this is a problem of lack of awareness which is grounded in an emotional detachment between the users and their digital data. In line with this, we will discuss preventative technologies for raising awareness and approaches for building a stronger connection between the users and their private information. Likewise, we encourage the inclusion of awareness mechanisms for providing better insights on the privacy policies of SNSs.

1 INTRODUCTION

In 1966, tobacco companies across the United States were affected by a law that later on changed the standards for the commercialization and distribution of cigarettes. For the first time in the history, a legislation requiring warnings about the risks associated with the consumption of tobacco was proposed by the U.S Congress (Hiilamo et al., 2014). Since then, the companies began fighting against Health Warning Labels (HWLs) in cigarette packs basically arguing that people already knew the hazards of smoking. Despite their efforts on blocking or weaken HWLs, nowadays many countries have included and implemented HWL in their legislations (Hiilamo et al., 2014).

Social Network Sites (SNSs) are spaces which are not free of privacy risks, and like in the case of cigarettes consumers, users of SNSs might have heard about some of these risks before or during their activity period (i.e. before or after opening an account in a SNS). While one might argue that the risks of disclosing personal or sensitive information in SNSs are not as severe as the risks of smoking, unwanted incidents such as job loss, reputation damage, or unjustified discrimination should not be neglected or disregarded. However, very little information (for not to say none) is provided by the SNSs about the potential risks of information sharing.

Privacy policies can be considered as an initial approach towards the information on potential privacy risks. However, these electronic documents are shown once to the users (when registering), and are hardly revised by them in the future. Moreover, privacy policies basically inform about which data is collected, how is processed, and under which conditions it is disclosed to third parties; without any emphasis on informing about potential risks. If we add to this that users are not strongly attached to their private information, then the chances of users regretting to have shared private information increases.

We believe that, like tobacco consumers, the users of SNSs should be empowered with information about the potential risks of information sharing. Moreover, we believe that awareness mechanisms can be a good alternative not only to inform the users about such risks, but also to create a stronger tie between them and their private information. In this paper we take a closer look at the privacy risks associated with user-generated content in SNSs in order to discuss possible solutions to this issue. More-

over, we provide arguments towards the use of adaptive preventative technologies to move towards a more privacy-aware social environment in SNSs.

The rest of the paper is organized as follows. In the next section, we discuss the motivation scenario and the paper's background. In Section 3, we analyse the role of privacy policies and media technologies on modulating users' perceptions towards their private data. Following, we discuss in Section 4 preventative technologies for the generation of awareness within SNSs. In Section 5, we analyse an approach for incorporating privacy heuristics derived from users' regrettable experiences into the design of preventative technologies. Thereafter, we discuss the advantages and drawbacks of this approach in Section 6. Finally, we conclude in Section 7 with an outlook and considerations for further research.

2 Motivation and Background

In 2018, the EU's new General Data Protection Regulation (GDPR) (Regulation, 2016) will come into force as the conclusion of a hectic debate which has involved academics, Internet service providers and international organizations across the world. For many, the Internet is considered an open platform for democratic participation which promotes freedom of expression and the right to information access. Therefore, is not surprising that the GDPR, and more specifically the Right to be Forgotten (RTBF), raises concerns related to abusive removal demands of user-generated content (e.g. public officers trying to suppress criminal records), and other issues about potential unjustified censorship[1]. This is a debate which mainly circles around the right to erase or de-list information put online by another Internet user. In this work, we do not aim to discuss this aspect of the GDPR. Instead, we look to resume the discussion to the information that users disclose about themselves in SNSs.

One of the critical concepts included in the GDPR is the one of "personal data". For instance, Article 4 says that information related to a "data subject" (i.e. an identified or identifiable natural person[2]), such as name, identification number, location, factors specific to his/her physical, physiological, genetic, mental, economic, cultural or social identity, should be considered as personal information and therefore require *unambiguous* consent to be processed. Likewise, Article 9 says that racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade-union membership should not be processed unless the data subject gives *explicit* consent. *Unambiguous* consent can be given through a conduct that clearly indicates that the data subject agrees with the proposed processing of his/her personal data (e.g. when telling the doctor about the medical ailment one is suffering while he/she enters notes in a computer system). On the other hand, *explicit* consent should be given trough an explicit action by the data subject. This is normally granted after the data subject clicks on 'Yes, I agree" in the *privacy policies* of the service provider.

[1]The work by Keller (Keller, 2017), offers a clarifying view on the RTBF and its hazards for freedom of expression and information rights on the Internet.

[2]An identifiable person is one who can be identified, directly or indirectly.

3 Data "Visceralization"

One of the main reasons for differentiating private data from general data are the risks associated with their inappropriate processing and public disclosure. Basically, the GDPR encloses an implicit warning message for data processors (the SNSs in our case) which is that they should safeguard data subjects from unwanted incidents (such as unjustified discrimination, political or religious persecution, or fraud) by treating carefully their personal data. Privacy policies, on the other hand, inform the users about which information will be collected, processed, used, disclosed and managed by the data processor. As one can observe, there is a semantic difference in the message of the GDPR and the one of privacy policies. Whereas the GDPR endows service providers with a better perception on the importance of the users' personal information, privacy policies do not provide cues to data subjects about the importance of their own personal information. Consequently, privacy policies in some point modulate users' perceived severity of privacy risks in SNSs.

Like privacy policies, information sharing interfaces of SNSs also play an important role in shaping our perceptions of information privacy (Stark, 2016; Díaz Ferreyra et al., 2017a). Such interfaces are the entry point of user-generated content which, in many cases, contains private information. However, since digital data is intangible, information sharing interfaces of SNSs regulate users' emotional perception and attachment towards their private information. Let us consider the following example: imagine that a stranger stops you in the street and asks you for your passport. It is quite unlikely that someone would grant this request in the real world. Moreover, this situation would normally come along with a *visceral reaction* (i.e. an instinctive gut-deep bodily response

like a burning sensation in the stomach) as consequence of this unexpected request. However, when this information is requested through the interfaces of a SNSs, such reactions do not seem to arise. Consequently, privacy policies and sharing interfaces are not succeeding in taking the users' emotional perception of their private data to the visceral level.

4 Privacy Awareness in SNSs

Like in the case of HWL for the commercialization of cigarettes, awareness mechanisms for SNSs can contribute to bridge the emotional gap between users and their digital data. In this section we discuss different preventative technologies oriented to generate awareness in online self-disclosure scenarios. This is, scenarios in which users intend to reveal their own private information in SNSs.

4.1 Preventative Technologies

Different preventative technologies have been proposed for mitigating the unwanted consequences of online self-disclosure (Calikli et al., 2016; Díaz Ferreyra et al., 2016; Fang and LeFevre, 2010; Wang et al., 2013; Ghazinour et al., 2013). One of the most representative of these approaches is the one by Wang et al. (Wang et al., 2013) consisting of three plugins for Facebook. These plugins called "privacy nudges" intervened when the user was about to post a message in his/her Facebook biography either (i) introducing a delay, (ii) providing visual cues about the audience of the post, or (iii) giving feedback about the meaning (positive or negative) of the post. However, since the feedback generated by the nudges was the same for every user of Facebook, they did not succeed on reaching high levels of acceptance. This is, some users liked them and others found them annoying. Consequently, this type of technology should provide adaptive feedback and awareness to their users in order to being widely adopted.

4.2 Adaptive Awareness

Adaptive preventative technologies seek to develop mechanisms capable to provide tailored feedback and awareness to their users. One of the preventative technologies which follow this direction is the one of Ziegeldorf et al., consisting in a framework of personalized privacy metrics for the generation of adaptive awareness (Ziegeldorf et al., 2015). This approach, called Comparison-based Privacy (CbP), consists of analysing different *comparison metrics* which are computed over the content being shared among different *comparison groups*. Basically, comparison groups consist of groups of people with which the user can intuitively relate to (e.g. family, friends and colleagues, users with the same profession or same age). Likewise, comparison metrics capture aspects of the sharing behaviour within a privacy group, such as the sentiment and the type of the content being shared. A user can choose for instance to compare the amount of hate speech in his/her posts against the one of people with his/her same profession. If this value exceeds a given threshold, then the system alerts the user. Thresholds can be set individually by users, or according to general profiles representing different privacy attitudes (e.g. unconcerned, pragmatist or fundamentalist). Approaches like this one overcome the engagement issue caused by generic warning messages of static approaches.

5 Visceral-Awareness Design

One of the key elements of HWLs in cigarettes packaging is that they include pictorial representations of the risks of tobacco consumption. This is done in order to make users perceive such risks in a more visceral way. In the case of online self-disclosure, regrettable experiences come along with visceral reactions from the users. This is, when a user lives an unwanted incident after disclosing personal data in SNSs, then a feeling of regret and repentance arises together with a visceral reaction. In this section, we discuss design principles introduced by Díaz Ferreyra et al. to include regrettable experiences into the design process of preventative technologies.

5.1 Privacy Heuristics

Díaz Ferreyra et al. propose to take into account regrettable self-disclosure experiences in order to endow preventative technologies with visceral-awareness principles (Díaz Ferreyra et al., 2017a). Basically, they suggest that privacy heuristics (best practices) can be derived from regrettable self-disclosure experiences and used thereafter to raise privacy awareness. For this, they introduce a Privacy Heuristics Derivation Method (PHeDer) for eliciting privacy best practices from user's regrettable experiences (Díaz Ferreyra et al., 2017a). The first step of this method, called *Regret Acknowledgement*, consists on gathering evidence about a regrettable experience and describe it in terms of: (i) the information that was disclosed (ii) the unintended audience it reached, and (iii) the unwanted incidents that lead the

user to a feeling of regret. The output of this step can be represented as in Fig. 1, where the user reported to have shared his/her political affiliation in a public post. Once the regrettable scenario is described, it is forwarded to the next step called *Regret Analysis*.

The *Regret Analysis* step consists of refining the scenario of Fig. 1 into privacy risks consisting of a 7-tuple of elements: a list of *personal attributes*, the *unintended audience*, the *unwanted incident*, the *frequency* of the unwanted incident, the *impact* of the unwanted incident, the *risk level*, and the user's *privacy attitude*. Whereas the personal attributes can be derived from articles 1 and 9 of the GDPR, frequency and impact of the unwanted incident (and consequently the risk level) can be expressed using nominal scales. Privacy attitudes are one of the *pragmatist* (medium privacy concern), *fundamentalist* (high privacy concern), or *unconcerned* (low privacy concern). For the example of Fig. 1, the output of this step would be risk([political_opinion], work_colleagues, wakeup_call, likely, major, very_high, pragmatist). This information if then forwarded to the third step of the method which is *Heuristic Design*.

USER'S POST

"Seriously? Trump became president? What is happening to the world!? #republicanssuck"

Actual Audience: PUBLIC.
Unintended Audience: The user's work colleagues.
Unwanted Incidents: Wake-up call from superior.

Figure 1: Example of self-disclosure scenario

The *Heuristic Design* step uses Constraint Based Modeling (CBM)(Mitrovic and Ohlsson, 2006) as the design principle for encoding the outcome of step 2 into a privacy heuristic. In CBM a constraint consists of a pair of *relevance* and *satisfaction* conditions, where each member of the pair can be seen as a set of features or properties that a disclosure scenario must satisfy. For the given example, the relevance conditions would be the existence of a political opinion inside a post, and the satisfaction condition would be not include the work colleagues in the post audience. Such constraints can be expressed using Horn clauses in Prolog and be included in the final step (*Constraint Integration*) in the Privacy Heuristics Data Base (PHDB) of an Instructional Awareness System (IAS) (Díaz Ferreyra et al., 2017a; Díaz Ferreyra et al., 2016).

5.2 Instructional Awareness

Basically, an IAS uses the heuristics inside a PHDB to detect potentially regrettable disclosures. Such heuristics are evaluated when a "post" event takes place in order to determine if the disclosure action can derive in a regrettable scenario for the user. This is, done first by evaluating the *relevance* condition of the heuristics. Let us consider a scenario where a user wants to disclose once again his/her *political affiliation* inside a public post. Let us also consider that the heuristic discussed in Section 5.1 is part of the PHDB. In this case, IAS will detect that the disclosure can lead to a potential regret, and therefore proceeds to raise a warning to the user.

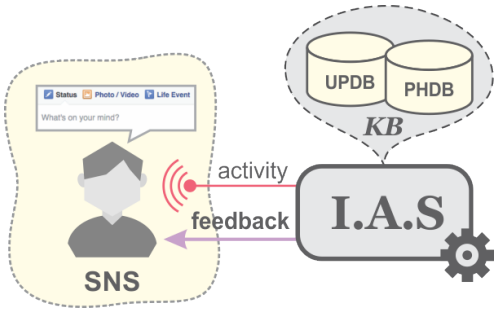


Figure 2: Instructional Awareness System (IAS)

In order to generate adaptive feedback, IAS takes into account adaptivity variables such as the user's privacy attitude, the number of times the user has ignored/accepted the warnings, and how often he/she discloses private information. This information is stored in a User Performance Data Base (UPDB) that, together with the PHDB, makes up IAS's Knowledge Base (KB). With the information stored in its KB, IAS can generate a message such as "Revealing your political affiliation to your work colleagues can bring you problems. Do you want some hints on how to protect your private data?" and recommend the user to restrict the post's audience (for instance to "friends only"). Since information about the risks is also kept in the KB, an IAS can also provide such additional information in the warning message.

6 Discussion

Following a similar approach to the HWLs in cigarettes packages, Díaz Ferreyra et al. propose to inform the users of SNSs about the risks of online self-disclosure though an IAS. For this, IAS requires risk knowledge which is stored in a PHDB and obtained through a privacy heuristics derivation method.

This method is an *offline* approach for eliciting privacy heuristics from regrettable online self-disclosure experiences. Basically, the input of the method are the experiences that users have reported themselves (to the development team of IAS for instance), or the outcome of an empirical research (e.g. questionnaires or face to face interviews). This approach is effective for building a baseline of heuristics prior to the execution of the system. However, eliciting new entries of the PHDB requires the execution of this process which can be expensive and inefficient in terms of the resources and time needed to conduct interviews and process the outcome of them. One way to overcome this issue is to consider deleted posts with private information as potential sources of heuristics (Díaz Ferreyra et al., 2017b). This is, to use such posts as the input of a machine learning engine for the automatic derivation of privacy heuristics at runtime. This way, the PHDB can be updated with new heuristics without having to execute offline iterations of the PHeDer method.

7 Outlook and Conclusion

Adaptive awareness technologies seem to be promising approaches for empowering the users of SNSs in making wiser and more informed decisions, as to protect them from the risks of over-sharing private information. We believe that this is not a minor issue that should be taken seriously into consideration by Internet service providers, multilateral organizations and policy makers. We have used the example of HWLs in cigarettes packages as a motivating scenario for working towards a more privacy aware social environment in SNSs. Certainly, this topic will be part of a in-depth and intense debate in the future. Therefore we hope this paper will offer a more clarifying view on this issue and serve as an instrument for the development of more effective solutions.

ACKNOWLEDGEMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

REFERENCES

Calikli, G., Law, M., Bandara, A. K., Russo, A., Dickens, L., Price, B. A., Stuart, A., Levine, M., and Nuseibeh, B. (2016). Privacy Dynamics: Learning Privacy Norms for Social Software. In *Proceedings of the 11th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, pages 47–56. ACM.

Díaz Ferreyra, N. E., Meis, R., and Heisel, M. (2017a). Online Self-disclosure: From Users' Regrets to Instructional Awareness. In *Proceedings of the IFIP International Cross-Domain Conference (CD-MAKE)*. Accapted for publication.

Díaz Ferreyra, N. E., Meis, R., and Heisel, M. (2017b). Towards an ILP Approach for Learning Privacy Heuristics From Users' Regrets. In *Proceedings of the 4th European Network Intelligence Conference (ENIC)*. Accapted for publication.

Díaz Ferreyra, N. E., Schäwel, J., Heisel, M., and Meske, C. (2016). Addressing Self-disclosure in Social Media: An Instructional Awareness Approach. In *Proceedings of the 2nd ACS/IEEE International Workshop on Online Social Networks Technologies (OSNT)*. ACS/IEEE.

Fang, L. and LeFevre, K. (2010). Privacy wizards for social networking sites. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 351–360, New York, NY, USA. ACM.

Ghazinour, K., Matwin, S., and Sokolova, M. (2013). YourPrivacyProtector: A Recommender System for Privacy Settings in Social Networks. *International Journal of Security, Privacy and Trust Management (IJSPTM)*, 2(4).

Hiilamo, H., Crosbie, E., and Glantz, S. A. (2014). The evolution of health warning labels on cigarette packs: the role of precedents, and tobacco industry strategies to block diffusion. *Tobacco control*, 23(1):e2–e2.

Keller, D. (2017). The right tools: Europe's intermediary liability laws and the 2016 general data protection regulation. Technical report, Stanford Law School Center for Internet and Society.

Mitrovic, A. and Ohlsson, S. (2006). Constraint-based knowledge representation for individualized instruction. *Computer Science and Information Systems (ComSIS) Journal*, 13(1).

Regulation, G. D. P. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union (OJ)*, 59:1–88.

Stark, L. (2016). The Emotional Context of Information Privacy. *The Information Society*, 32(1):14–27.

Wang, Y., Leon, P. G., Scott, K., Chen, X., Acquisti, A., and Cranor, L. F. (2013). Privacy Nudges for Social Media: An Exploratory Facebook Study. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 763–770. ACM.

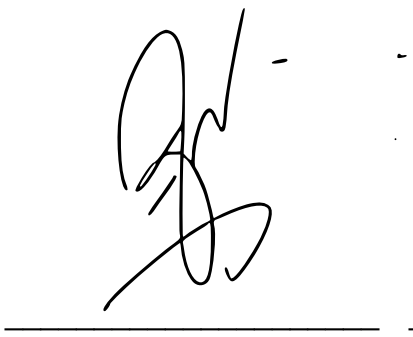
Ziegeldorf, J. H., Henze, M., Hummen, R., and Wehrle, K. (2015). Comparison-based privacy: nudging privacy in social media (position paper). In *International Workshop on Data Privacy Management*, pages 226–234. Springer.

DECLARATION OF INDIVIDUAL CONTRIBUTIONS

Publication title: At Your Own Risk: Shaping Privacy Heuristics for Online Self-disclosure

Reference item: N. E. Díaz Ferreyra, R. Meis, and M. Heisel, "At Your Own Risk: Shaping Privacy Heuristics for Online Self-disclosure," in *16th Annual Conference on Privacy, Security and Trust (PST).* IEEE, September 2018, pp. 1–10. doi: 10.1109/PST.2018.8514186

Author	Contribution	%
Nicolás E. Díaz Ferreyra	- Conceptualisation of the approach. - Planification of the work. - Draft of the manuscript.	90%
Rene Meis	- Supervision and advice.	5%
Maritta Heisel	- Supervision and advice.	5%



Nicolás E. Díaz Ferreyra

Rene Meis

Maritta Heisel

At Your Own Risk: Shaping Privacy Heuristics for Online Self-disclosure

Nicolás E. Díaz Ferreyra*, Rene Meis†, Maritta Heisel‡
University of Duisburg-Essen, Germany
RTG User-Centred Social Media
https://www.ucsm.info/
{nicolas.diaz-fereyra*, rene.meis†, maritta.heisel‡}@uni-due.de

***Abstract*—Revealing private and sensitive information on Social Network Sites (SNSs) like Facebook is a common practice which sometimes results in unwanted incidents for the users. One approach for helping users to avoid regrettable scenarios is through awareness mechanisms which inform a priori about the potential privacy risks of a self-disclosure act. Privacy heuristics are instruments which describe recurrent regrettable scenarios and can support the generation of privacy awareness. One important component of a heuristic is the group of people who should not access specific private information under a certain privacy risk. However, specifying an exhaustive list of unwanted recipients for a given regrettable scenario can be a tedious task which necessarily demands the user's intervention. In this paper, we introduce an approach based on decision trees to instantiate the audience component of privacy heuristics with minor intervention from the users. We introduce Disclosure-Acceptance Trees, a data structure representative of the audience component of a heuristic and describe a method for their generation out of user-centred privacy preferences.**

***Index Terms*—social network sites, adaptive privacy, awareness, heuristics, risk analysis, human-computer interaction**

I. Introduction

Disclosing personal information to others is a common practice which is key for the development and maintenance of social relationships. Social Network Sites (SNSs) like Facebook or Twitter encourage and foster such process through a set of features which allow their users to share information with large and diverse audiences. Whereas this contributes to develop and strengthen the social links between people, it can also result into regrettable and negative experiences for the users when pieces of private information reach an unintended audience [1].

Mapping disclosures to the right audience can be a hard work for a regular user of a SNS [2, 3]. The increasing size of social connections over time and the lack of a strong criterion for segmenting the audience make privacy a hard task to carry forward. In order to support users in keeping their content away from the wrong audience, privacy scholars have proposed a wide range of *preventative technologies* [4–6] which generate awareness on the content being disclosed [7]. Basically, these technologies provide cues about the semantics of the user's posts (i.e. if contains private information or not) and in some cases suggest a course of action to overcome potential privacy issues. However, the risk factor of information disclosure is hardly taken into consideration in the design of preventative technologies and thus in their recommendations.

978-1-5386-7493-2/18/$31.00

In order to find the right audience for a post, the potential privacy risks of its disclosure should be evaluated. In other words, to disclose or not a post is a decision which is often grounded in the acknowledgement of its potential negative consequences [8]. For instance, a user Alice could decide to exclude her work colleagues from the audience of a post which contains a negative comment about her workplace based on the notion that this can bring her problems with her employer in the future. This is because risks are connected with *visceral* emotions (e.g. fear, anxiety or distress) which help us to reflect over the potential consequences of an action [9, 10]. Therefore, knowledge about the risks of recurrent self-disclosure scenarios like this one can serve preventative technologies in the generation of privacy awareness.

Díaz Ferreyra et al. [11] propose to encode regrettable self-disclosure scenarios as *privacy heuristics* in order to evaluate the risks of sharing private information in SNSs. Basically, heuristics are a representation of empirical evidence about recurrent regrettable self-disclosure scenarios and hold knowledge about the risks of disclosing particular pieces of private information to specific audiences. For instance, a heuristic can describe the scenario in which the user gets a wakeup call from a superior after posting a negative comment about his/her workplace. Consequently, preventative technologies can use the knowledge inside privacy heuristics to evaluate if a particular post can lead the user to a well-known regrettable situation.

The *audience* is one of the most important components of a heuristic since it represents the group of people who should not access particular information under certain privacy risks. Nevertheless, this component is often hard to specify since it requires to create and maintain exhaustive *allow/deny* lists of information recipients. Moreover, since this component varies from individual to individual, the users themselves are the ones who should carry out this tedious task. In this paper we introduce an approach based on decision trees which allows to determine the audience component of a heuristic with minor intervention from the user. Since these data structures can be learned out of a training set, users are only requested to provide a reduced number of examples in order to create a decision tree representative of an audience. Moreover, this

allows to incorporate user-centred privacy preferences in the design and development of preventative technologies.

The rest of the paper is organized as follows. In the next section we discuss related work on privacy awareness and on the automatic generation of access-control policies. In Section III we introduce the concepts of privacy heuristics and regrettable scenarios and analyse their impact on the design of preventative technologies. Next, Section IV introduces the concept of Disclosure-Acceptance Trees, a supervised-learning method for their construction, and an application example. Section V discusses the impact of our approach for the development of public policies. Finally, in Section VI we conclude and discuss future work.

II. Related Work

Privacy scholars have proposed several strategies for generating privacy awareness in SNSs. Among them, *nudging* is a well-known approach in which soft paternalistic interventions (i.e. information and guidance) attempt to influence users' decisions towards safer and better choices without imposing a particular course of action [12]. In line with this approach, Wang et al. [13] developed three nudging strategies for Facebook consisting of (a) delaying the time before a post appears on the user's profile, (b) displaying visual cues about the post's audience, and (c) analysing and displaying the post's sentiment to the user. These nudges intervened when users attempted to post a message on Facebook allowing them to reconsider their disclosures and avoid regrettable scenarios. Whereas this is an interesting approach for generating awareness, it can nevertheless be too invasive for the users when their particular privacy expectations are not taken into account. In other words, nudges can fail on engaging with those users who are less privacy-concerned or hold higher levels of privacy literacy when the frequency and intensity of the warnings is not regulated [14].

Another approach for privacy protection in SNSs are access-control policies (i.e. privacy settings) used to specify which information ought to be accessed by who in a particular context. Several approaches have been proposed for the automatic generation of such policies with regards to user-generated content in SNSs [4, 5, 15]. Many of these approaches analyse *what* has been shared with *whom* in the past in order to recommend policies aligned with the users' privacy practices. For instance, Misra and Such [15] developed a predictive model which considers previous sharing actions of the users on Facebook for the generation of access-control policies. In this case, the generated policy prescribes which type of content should be shared with whom depending on the type and strength of the online relationships a user maintains with his/her Facebook friends. Although machine learning approaches like this one show acceptable levels of accuracy, these methods still rely on the assumption that users have shared their content with the right audience in the past. Consequently, inferred policies do not always meet the users' privacy preferences and expectations especially for those users who show a great variation in their access-control decisions [16].

In order to reflect the true privacy preferences of the users, Fang and LeFevre [5] suggest an *active learning* approach in which users are interactively queried about their privacy preferences. In their approach, the privacy preferences of a user (i.e. the user's willingness or unwillingness to share information with each of his/her friends) are elicited by asking him/her to assign privacy labels to a representative sample of friends. These privacy preferences are then used together with a set of features (i.e. community structure and other information available inside the user's profile) to build a privacy-preference prediction model. We will follow this active learning direction for learning the audience component of privacy heuristics. Since heuristics introduce explicitly the risks of self-disclosure, they do not only allow to answer the question "*what* is right to be disclosed to *whom*?", but also "*what-if* private/sensitive information is disclosed?". In the next section we discuss in detail the use of privacy heuristics and their advantages for the development of preventative technologies like nudges.

III. Preeliminaries

In this section we introduce three basic design principles for the development of preventative technologies and elaborate the concept of privacy heuristics around them. Likewise, we discuss the role of regrettable online experiences for shaping privacy heuristics together with methods proposed for their derivation.

A. Awareness Design Principles

As mentioned previously, preventative technologies can fail in engaging with their users mainly for not taking into account their literacy level and personality traits (e.g. privacy attitudes). However, there are other aspects regarding the nature of private digital data and self-efficacy (i.e. how well a user can execute actions required to deal with risky privacy scenarios) that should be considered in the design of these technologies [7]. In order to guide preventative technologies towards higher levels of engagement, we introduce three design principles that should be taken into consideration for their development. Based on relevant findings in the areas of privacy awareness and online self-disclosure, we suggest that preventative technologies should incorporate *adaptive*, *visceral* and *supportive* principles to their design.

The *adaptive* principle refers to the fact that users do not have the same expectations and behaviour regarding their own privacy. Some users are more willing to disclose private information online without much concern about the consequences, and others rather keep such information away from unwanted recipients [11]. If an unconcerned user gets too many warning messages when he/she wants to share something on the Internet, then the system would probably fail on its propose for being too annoying. Consequently, preventative technologies should generate warnings aligned with the users'

privacy expectations in order to engage them in a continuous learning process.

The *visceral* principle is related with the role of technology in shaping our perception of information privacy. Since private digital data is intangible, it is perceived only through the interfaces and physical materials of media technologies. Consequently, such technologies modulate in a certain way users' emotional perception and attachment towards their private information [17]. However, media technologies are not succeeding in taking such emotional perception to the visceral level. This is, making the tie between users' feelings and data visible, tangible and emotionally appreciable so they can perceive (in a visceral way) the impact of their disclosures [17]. Consequently, preventative technologies should generate a visceral connection between users and their disclosures.

Finally, preventative technologies should be *supportive* in order to help users to overcome potential privacy issues. This is, warning messages should not only inform about the sensitivity of the content being disclosed, but they should also recommend *heuristics* that the users can put into practice in order to preserve their privacy [11]. Access-control policies can serve this aspect since they help the user to make decisions about the post's audience. However, such decisions should not be only based on the users' previous privacy practices but also on some segmentation criterion.

B. Regrettable Experiences

A good criterion for audience segmentation (and therefore for privacy heuristics) can be found in the regrettable self-disclosure experiences of the users. Basically, a self-disclosure regret is a feeling of sadness, repentance or disappointment which occurs when a piece of sensitive information reaches an unintended audience and derives in an unwanted incident [1, 11]. For example, Fig. 1 describes a scenario in which a user regrets to have shared a post with a negative comment about his/her workplace. In this scenario, the comment reached the user's work colleagues (i.e. the unintended audience) and derived in a wake-up call from the superior, together with a bad image (i.e. the unwanted incidents). As one can see, these negative experiences enclose knowledge about in which cases particular pieces of personal information should be kept away from specific groups of people. Moreover, since regrets often come along with a *visceral reaction*, they are good resources for the generation of visceral privacy awareness.

Under this approach, privacy awareness can be raised by conducting a risk analysis over user generated content. This is, using regrets as sources of knowledge, one can analyse the *privacy risks* (i.e.*consequence* and the *frequency* of the unwanted incidents) associated with the disclosure of particular content to a particular audience. Risks, consequence and frequency, can be expressed in a nominal scale. In this sense, a consequence is a value on an impact scale such as insignificant, minor, moderate, major or catastrophic. Likewise, the frequency is a value on a likelihood scale such as rare, unlikely, possible, likely and certain. Finally, a risk is a value obtained from the frequency and consequence of the unwanted incident and expressed in a scale such as very low, low, high and very high.

USER'S POST

"A typical day at the office. Lots of complaints and bad mood. Cannot wait for the day to be over...!"

Actual Audience: PUBLIC.
Unintended Audience: The user's work colleagues, or superior.
Unwanted Incidents: Wake-up call from superior; bad image; job loss.

Fig. 1. Example of self-disclosure scenario

C. Privacy Heuristics Derivation

A privacy heuristic is basically a tuple $<$*PAs, Audience, Risk*$>$ where *PAs* is a set of private attributes, *Audience* is a collection of recipients (e.g. Facebook *friends*), and *Risks* corresponds to the frequency and consequence of an unwanted incident. A heuristic models the relation between these three elements as follows: *"The privacy Risk (i.e. consequence and frequency of an unwanted incident) associated with the disclosure of a set of PAs to an unintended Audience"*. So far, two methods for the derivation of privacy heuristics have been proposed. The first one, called Privacy Heuristics Derivation Method (PHeDer), consists of gathering empirical evidence of regrettable online experiences using questionnaires and face-to-face interviews [11]. Using this evidence as input, the PHeDer method proceeds to the derivation of the heuristics by analysing the regrettable scenarios in terms of the unintended audience, the information disclosed and the associated privacy risks. At the end of the method, such information is represented by a set of heuristics encoded as Horn Clauses in Prolog. The second method uses Inductive Logic Programming (ILP) as the approach for deriving the heuristics [18]. Basically, ILP employs techniques from machine learning and logic programming to infer a hypothesis from a set of positive and negative examples and some background knowledge [19]. In this case, deleted posts with sensitive information are considered as explicit manifestations of the users' regrets and used as the training set. Since deleted posts do not include information about the privacy risks and the unintended audience, this is asked to the users though a special interface. Such information is then used by an ILP engine to infer a complete and consistent hypothesis which in this case is a privacy heuristic.

IV. Shaping Privacy Heuristics

Up to now, the methods that have been proposed for the derivation of privacy heuristics do not model their audience component explicitly. Instead, they express audiences using abstract labels of common social identity circles (e.g. family, work colleagues, university friends). This is certainly a

limitation of these methods since labels do not provide any cues about which friends should be included inside those circles. Therefore, the audience component as such (i.e. as an abstract label) cannot serve in the automatic evaluation of future disclosures. An approach to overcome this issue, could be a decision-support model generated from a sample set of the people that the user is willing to share his/her information with.

Decision trees (DTs) create a model that predicts the *value* of a target *variable* by learning simple decision rules inferred from a training set (i.e. observed features) [20, 21]. DTs are at their heart a fairly simple type of classifier which are used to sort previously unseen examples. Therefore, one could generate a DT from a *reduced sample* of the people that the user is willing to share private information with. This is, build a DT out of a training set consisting of the people a user is willing to share a particular set of *PAs* under a certain privacy *Risk*. Thus, the *Audience* element of a heuristic would be represented by a Disclosure-Acceptance tree (DAT) that is capable to classify the recipients of PAs in function of the potential privacy risks. Following, a description of the development process of DATs for privacy heuristics is presented.

A. DAT Learning Process

The learning process of a DAT consists of four sequential steps which are *Content Analysis*, *Heuristics Matching*, *Privacy Preferences Elicitation*, and *DAT Generation*. As depicted in Fig. 2, each stage of the method draws on different external inputs and generates the outputs for the next step. The final output of the method is a DAT which models the *Audience* component of a privacy heuristic.

Step 1: Content Analysis This first step of the method is triggered by a *Post* event. Such event is nothing but an action performed by the user, in which he/she attempts to disclose some information (private or not) inside a SNS. In this step, the content of the message is analysed for the identification of private information. In order to conduct this task, one must define what information should be deemed for privacy analysis through a taxonomy of personal attributes [11]. However, what information is ought to be or not considered for privacy analysis depends largely on the purpose and the context in which such analysis is conducted. In this case, the risks of online self-disclosure define the context and purpose for our analysis. Therefore, a taxonomy of attributes must cover those aspects of the users' personal information associated with-well known risks of online self-disclosure. For instance, if a *job loss* incident is the consequence of revealing a negative comment about one's workplace, then the taxonomy should cover the attributes *work place* and *negative sentiment*.

In [11] Díaz Ferreyra et al. introduce a taxonomy consisting of a set of identifiable attributes of private and sensitive nature called Surveillance Attributes (SAs). Such SAs are organized around a number of high level categories called "self-disclosure dimensions" which classify them into *demographics, sexual profile, political attitudes, religious beliefs, health factors and condition, location, administrative, contact, and sentiment* (Table I). Although this taxonomy is not explicitly aligned with the risks of online self-disclosure, it has a strong correlation with regrettable scenarios reported by users of SNSs in a study by Wang et al. [1]. Furthermore, it provides an intuitive representation of different aspects of the users' private information in SNSs. We will adopt this taxonomy in this step and use it to identify relevant private information (i.e. SAs) inside the user's post (Fig. 4). For the example of Fig. 1, such attributes are *employment status*, *work location*, together with a *negative sentiment*; which correspond to the dimensions *demographics*, *location* and *sentiment* respectively. This step can be automated using Natural Language Processing techniques and methods[1].

Dimension	**Surveillance Attributes**
Demographics	Age, Gender, Nationality, Racial origin, Ethnicity, Literacy level, Employment status, Income level, Family status
Sexual Profile	Sexual preference
Political Attitudes	Supported party, Political ideology
Religious Beliefs	Supported religion
Health Factors and Condition	Smoking, Alcohol drinking, Drug use, Chronic diseases, Disabilities, Other health factors
Location	Home location, Work location, Favorite places, Visited places
Administrative	Personal Identification Number
Contact	Email address, Phone number
Sentiment	Negative, Neutral, Positive

TABLE I
THE "SELF-DISCLOSURE" DIMENSIONS.

Step 2: Heuristics Matching We will assume that the awareness system under development (i.e. under audience-learning phase) is endowed with a Privacy Heuristics Data Base (PHDB) used to detect risky self-disclosure scenarios. The heuristics inside this PHDB can be of two types: *baseline* or *personalized*. As shown in Fig. 3, *baseline heuristics* (BH_i in Fig. 4) are generic in the sense that they are only specified in terms of an unwanted incident, its frequency, and the disclosed SAs. Consequently, the DAT which represents the audience component of the heuristic has not been yet specified. On the other hand, *personalized heuristics* (PH_j in Fig. 4) have their audience component instantiated according to the user's privacy preferences, and the corresponding privacy risks. Thus, *baseline* heuristics are *personalized* as consequence of the process described in Fig. 2.

Let us assume that our PHDB counts with an initial set of baseline heuristics that have been generated following the methods described in Section III-C. Then, this step consists of selecting those *baseline heuristics* whose SAs match the

[1]Hoang et al. for instance, applied different machine learning algorithms for the automatic extraction of location information using a corpus of Tweets [22]

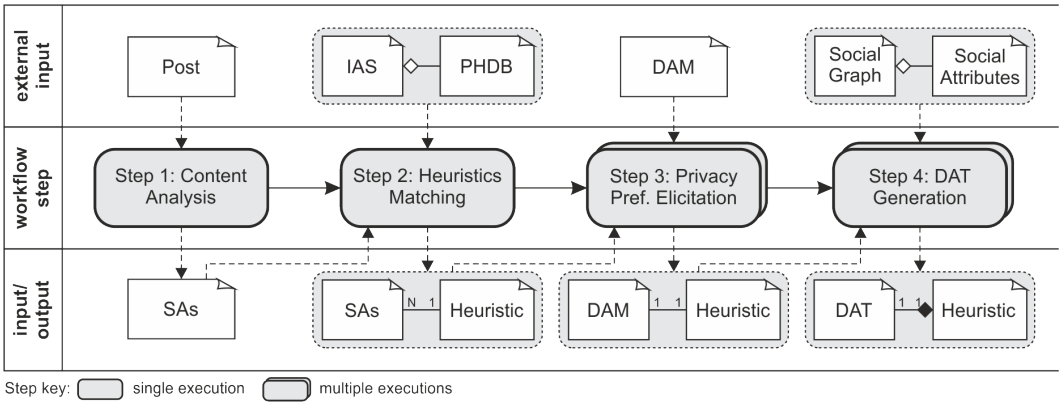


Fig. 2. DAT Learning Process

ones that the user intends to disclose in his/her post. That is: being $BH_i.SAs$ the SAs of a baseline heuristic and $Post.SAs$ the SAs inside the user's post, then BH_i matches the $Post$ when $BH_i.SAs \subseteq Post.SAs$ (as shown in Fig. 5). For the example of Fig. 1, the relevant baseline heuristics (i.e. the ones that match) are the ones whose SAs correspond to *employment status*, *work location*, and *negative sentiment*.

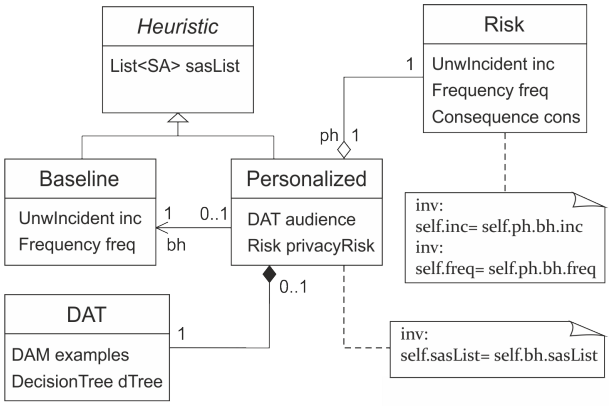


Fig. 3. Types of privacy heuristics

Step 3: Privacy Preferences Elicitation In order to generate the DAT of a heuristic, it is necessary to count with a training set of examples. As explained previously, this training set consists of a sample of friends that the user is willing to share the SAs in the post under certain privacy risks. For the scenario described in Fig. 1, such sample should be composed of people to whom the user agrees to disclose the SAs *employment status*, *work location*, and *negative sentiment* under the risk of losing his/her job. Since this information can only be provided by the user him of herself, an instrument for the elicitation of these privacy preferences must be provided. One way to retrieve such information is through a Disclosure-Acceptance Matrix (DAM) like the one of Table II. Each row of the DAM is a tuple $< Friend, Allow >$ where $Friend$ is someone from the user's friend-list, and $Allow$ is a yes/no value. Consequently, each matched heuristic in Step 2 requires a DAM for the generation of its corresponding DAT. In this step, an empty DAM is shown to the user, so he/she can give input about the trusted audience of the disclosed SAs. As one can observe in Fig. 4, the third component of a baseline heuristic BH_i is expressed as $[?, Unwi, Freq]$ where $Unwi$ corresponds to the unwanted incident, and $Freq$ to its frequency. The question mark $?$ refers to the *consequence* of the unwanted incident which (like the audience) is unknown by hand. Therefore, the user must also provide input about the perceived severity of the unwanted incident (i.e. the risk's consequence). As mentioned in Section III-B, this can be done by selecting a value from an impact scale such as *insignificant*, *minor*, *moderate*, *major*, or *catastrophic*.

Step 4: DAT Generation A DAT is a DT which classifies an audience instance (i.e. a Friend) into *trusted* or *not-trusted*

Fig. 4. Content Analysis and Heuristics Matching

Trusted Audience			
Friend	**Yes**	**No**	*example*
Alice		✗	e_1
Bob		✗	e_2
Kate		✗	e_3
John		✗	e_4
Susan	✓		e_5
Bill	✓		e_6
Robin		✗	e_7
James	✓		e_8
Sarah		✗	e_9
Jane		✗	e_{10}
Bridget		✗	e_{11}
Amanda		✗	e_{12}
Sam	✓		e_{13}
Marc	✓		e_{14}

Unwanted Incident: **Job Loss**

Consequence Level: insignificant | minor | moderate | major | **catastrophic**

TABLE II
DAM AND CONSEQUENCE LEVEL ELICITATION FOR "JOB LOSS"

by applying a number of conditional tests over a set of *social attributes* such as *age*, *gender*, *interests*, or *education*. Each test is contained in an internal node of the DAT together with a set of child nodes which represent the possible answers to the test. For instance, a node containing a test over the *Gender* attribute, will contain two child nodes for the values *Male* and *Female* respectively. Each child node is recursively defined as a node with a test over another social attribute and its corresponding list of nodes for each possible answer. The nodes of a DAT form a tree hierarchy of conditional tests over social attributes. An audience instance is sorted into a *class* (in this case *trusted* or *not-trusted*) by following the answers of each test from the root node (i.e. a topmost node) to a leaf node (i.e. a node without children). The instance is then assigned to the class that has been associated with the leaf it reaches. In general, DTs are incrementally generated by adding nodes which split the training set (in our case the DAM) into smaller and more homogeneous subsets [20, 21]. In the following section, we illustrate the learning process of a DAT with an example.

B. Application Example

Let us consider the DAM of Table 1 for the generation of a DAT, together with the social attributes *workplace*, *gender* and *location*. Let us assume that workplace can take the value *Company* when a friend works at the same company as the user does, *University* when a friend attends to the same university as the user does, and *Other* for any other case. Likewise, gender can be *Male* or *Female*, and location adopts the values $> 50\ km$ when a friend is based more than 50 km away from the user's location, or $\leq 50\ km$ in the opposite case. Table III expands the information contained in the DAM with these social attributes and their corresponding values. Each tuple of Table III is a training example used later on for the generation of the DAT.

Ex.	Friend	Workplace	Gender	Location	Trust
e_1	Alice	Company	Female	$>$ 50 km	No
e_2	Bob	Company	Male	$\leq$ 50 km	No
e_3	Kate	Company	Female	$\leq$ 50 km	No
e_4	John	Company	Male	$>$ 50 km	No
e_5	Susan	Other	Female	$>$ 50 km	Yes
e_6	Bill	Other	Male	$>$ 50 km	Yes
e_7	Robin	Other	Male	$\leq$ 50 km	No
e_8	James	Other	Male	$>$ 50 km	Yes
e_9	Sarah	Other	Female	$\leq$ 50 km	No
e_{10}	Jane	University	Female	$>$ 50 km	No
e_{11}	Bridget	University	Female	$\leq$ 50 km	No
e_{12}	Amanda	University	Female	$>$ 50 km	No
e_{13}	Sam	University	Male	$>$ 50 km	Yes
e_{14}	Marc	University	Male	$\leq$ 50 km	Yes

TABLE III
EXAMPLE SET FOR THE GENERATION OF A DAT

Algorithm 1 is an adaptation of the well-known ID3 algorithm used for the generation of DTs [20, 21]. The function *GenerateDAT* takes a *DAT* data structure, together with a list of *Attributes*, and a list of *Examples* as inputs. Such parameters represent the node which is under development (*currentNode*), a list with the social attributes to be considered (*attList*), and a list of the training examples (*exSet*) respectively. As mentioned before, a *DAT* is a recursive data structure containing a decision attribute *decisionAtt*, and a list of child nodes. The very first run of the algorithm starts with an empty DAT node, the complete list of attributes to be tested, and the full training set. In our example, the initial *attList* corresponds to $[workplace,\ gender,\ location]$ and the *exSet* to $[e_1, e_2, ..., e_{14}]$.

The *GenerateDAT* function starts by selecting which attribute best splits the *exSet* into nearly homogeneous/pure subsets of examples. This is, subsets where most of the example instances are labelled with one of the possible values of a *target attribute*. Since in our case the target attribute is *Trust*, the subsets generated after the split should be mainly composed by examples labelled with either *Yes* (i.e. trusted) or *No* (i.e. non-trusted) values. For this, a metric for measuring the *impurity* of the subsets must be defined. One of the most common measures used when constructing DTs is the *entropy* [20, 21]. The entropy is computed as $H(S) = \sum_{x \in X} -p(x).\log_2 p(x)$, where S is the example set being analysed, X a set of classification values, and $p(x)$ the fraction of the items in S which correspond to a classification value x. In our case, S corresponds to the *exSet* and X is composed by the classes *Yes/trusted* and *No/not-trusted*. When all the elements inside the set belong to the same class (i.e. a "pure" set of elements), the entropy value is 0. Conversely,

when the elements are assigned to different classes, the entropy value approximates to 1. Therefore, the best attribute for splitting a set of examples is the one which *minimizes* the entropy after the split.

Algorithm 1 DAT generation pseudo-code

```
1: function GENERATEDAT(var:DAT currNode,
   List<Attribute> attList, List<Example> exSet)
2:     Attribute bestAtt :=
           SelectBestAtt(attList, exSet);
3:     currNode.decisionAtt := bestAtt;
4:     attList.remove(bestAtt);
5:     for each Value val ∈
        currNode.decisionAtt.values do
6:         List<Example> exSubset :=
               ExSplit(exSet, bestAtt, val);
7:         DAT childNode := new DAT();
8:         if IsPure(exSubset) then
9:             Label lbl := GetLabel(exSubset);
10:            DATLeaf leafNode :=
                   new DATLeaf(childNode, lbl);
11:            currNode.add(leafNode);
12:        else
13:            currNode.add(childNode);
14:            GenerateDAT(childNode,
                   attList, exSubset);
15:        end if
16:    end for
17:    return
18: end function
```

The *information gain* measures the expected change in the entropy caused by partitioning the example set with regard to a certain attribute [20, 21]. The information gain of an attribute A relative to a collection of examples S is defined as $G(S, A) = H(S) - \sum_{v \in V(A)} \frac{S_v}{S}.H(S_v)$, where $V(A)$ are the possible values of an attribute A, and S_v is the subset of S for which the attribute A has value v (i.e. $S_v = \{s \in S \mid A(s) = v\}$). For our case, S corresponds to *exSet* and V to *attList*. As one can see, the first term of the equation corresponds to the entropy of the original set, whereas the second term is the expected value of the entropy after S is partitioned using the attribute A. To illustrate this concept, let us compute the information gain when $A = workplace$ and $S = exSet = [e_1, e_2, ..., e_{14}]$:

$$H(S) = -\frac{9}{14}.\log_2\frac{9}{14} - \frac{5}{14}.\log_2\frac{5}{14} = 0.94$$

$$H(S_{company}) = -\frac{4}{4}.\log_2\frac{4}{4} - \frac{0}{4}.\log_2\frac{0}{4} = 0.00$$

$$H(S_{university}) = -\frac{3}{5}.\log_2\frac{3}{5} - \frac{2}{5}.\log_2\frac{2}{5} = 0.97$$

$$H(S_{other}) = -\frac{3}{5}.\log_2\frac{3}{5} - \frac{2}{5}.\log_2\frac{2}{5} = 0.97$$

$$\begin{aligned} G(S, workplace) &= H(S) - \frac{4}{14}.H(S_{company}) \\ &\quad - \frac{5}{14}.H(S_{university}) - \frac{5}{14}.H(S_{other}) \\ &= 0.94 - \frac{4}{14}.0.00 - \frac{5}{14}.0.97 - \frac{5}{14}.0.97 \\ &= 0.24 \end{aligned}$$

The function *SelectBestAtt* returns the attribute in *attList* with the highest information gain with regards to *exSet* (Line 2). Since $G(S, gender) = 0.15$ and $G(S, location) = 0.09$, the best decision attribute is *workplace*. Consequently, this attribute is assigned as the decision attribute of the node (i.e. *currNode.decisionAtt*) and removed from *attList* (Lines 3 and 4).

Once the decision attribute for the current node is selected, we proceed to create branches below this node for each of its possible decision values (Line 5). For this, we split the *exSet* into subsets of instances associated with the same decision value (Line 6). Since *workplace* has three possible values, *exSet* is split into three subsets $[e_1, e_2, e_3, e_4]$, $[e_{10}, e_{11}, e_{12}, e_{13}, e_{14}]$, $[e_5, e_6, e_7, e_8, e_9]$ for the values *company*, *university*, and *other*, respectively. This process is executed iteratively, so that in each iteration the variable *exSubset* contains one of these subsets. Likewise, a new *childNode* is created on each iteration for each decision value and its respective *exSubset* (Line 7).

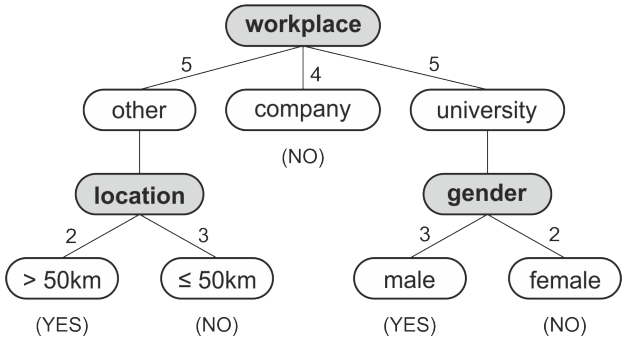


Fig. 5. Resulting DAT

The process of selecting a new decision attribute and partitioning the training examples is now repeated for each non-terminal node. This is, only for the child nodes whose *exSubset* is not pure. In order to determine if a *childNode* should be further split, the function *IsPure* computes the entropy of its corresponding *exSubset* returning *true* when is pure, and *false* when not (Line 8). Since all the instances of the *exSubset* corresponding to *company* are labelled as *NO*, the entropy is 0.00 and no further split is necessary (Lines 9-11). Therefore, this node corresponds to a leaf node in the *DAT* labelled with the value *NO*. Conversely, since the entropy values for

other and *university* is 0.97, a further split is required in both cases. Therefore, a recursive call to *GenerateDAT* is made with the corresponding *childNode*, its associated *exSubset*, and the *attList* as parameters (Lines 13 and 14). At the end of the process the *exSet* is split into a number of pure *exSubsets* which correspond to the leafs of the DAT. The final DAT learned by the Algorithm 1 from the examples of Table III is shown in Fig. 5.

C. Overfitting and Validation

A crucial aspect when applying DTs is limiting the complexity (i.e. the size) of the learned trees so they are capable to generalize examples beyond the ones of the training set [20]. Several alternatives have been proposed to deal with this issue, but basically they can be resumed into (a) approaches that stop growing the tree earlier before it starts *overfitting* (i.e. before it stats classifying too closely or perfectly the training data), and (b) approaches that allow the tree to overfit the data until no leaf can be further subdivided and then post-prune the tree [20, 21]. Due to the difficulty in the first approach of estimating precisely when to stop growing the DT, the second approach of post-pruning is the most commonly adopted in practice. Post-pruning can be achieved by collapsing internal nodes into leafs if this reduces the classification error. Other approaches remove nodes in an attempt to explicitly balance the complexity of the tree when it fits to the training data.

Another important aspect when constructing DTs is to determine if the generated tree generalizes beyond the training examples [20]. An approach often used for evaluating the performance of learning algorithms is *k*-fold cross-validation [20, 23]. Basically, this statistical method consists of dividing the data into *k* equally (or nearly equally) sized segments or *folds*. Subsequently, *k* iterations of *training* and *validation* are performed so that within each iteration a different fold is used for validation, while the remaining *k*-1 folds are used for training the model. This guarantees that all the available data has the chance to be validated against the model and the result is independent of the training/validation split. The cross-averaged accuracy of the *k* generated models during this process is then an estimate of the accuracy of a model trained using the full data set. In other words, this method allows to use all the data for estimating the accuracy of a model which is then built out of the same full data set. Consequently, with this method no data is left out for either training the DT or estimating its performance.

D. Strengths and Limitations

Using DTs for learning the audience component of a heuristic has certain advantages over other machine learning approaches. First of all, DTs are easier to code, manipulate and explain than other techniques such as Artificial Neural Networks or Support Vector Machines [20, 24]. Furthermore, DTs can be converted to a set of rules in order to make their representation more comprehensible. Another advantage is that they use a *white box* model, meaning that it is possible to closely examine their structure and understand the values of their output [24]. This is important in order to evaluate if the model is in compliance with the users' privacy preferences and make corrections when is necessary. Finally, DTs are capable to handle data sets that might have errors, and once constructed they can classify items quickly [20].

Despite the benefits we have just mentioned, there are certain limitations not only with regard to DTs, but also related to the overall approach we have introduced in this paper. First, DTs are sensible to changes in the input data [20]. This means that small changes in the examples used for their construction can lead to large changes in their structure and therefore in their predictions. Therefore, they require a certain precision from the users' input meaning that they should not be too ambiguous in their DAM choices. This leads to another issue which is related to the number of examples required for a good prediction. Although in this case the rule is "more is better", the selection of examples demands a cognitive effort from the user. This is something that should be considered for evaluation purposes and also as a point of improvement.

V. Discussion

Whether consciously or unconsciously, we interact with risky situations in our daily lives. From jaywalking or smoking cigarettes, people evaluate the potential consequences of risky actions on a daily basis. Likewise, risk-awareness strategies are used every day to inform people about the risks of engaging with certain activities or consuming products or services. For instance, Health Warning Labels (HWLs) have become a standard for the communication of the risks of smoking and are required for the commercialization of cigarette packages in many countries [25]. However, when it comes to SNSs, users are not given much information about the privacy risks of online interaction. Moreover, when users give their consent for data collection and processing (i.e. when they accept the *privacy policy*), they receive very little (for not saying none) information about such risks [7]. This lack of information modulates the perceived severity of privacy risks in favour of information disclosure and, consequently, in benefit of the service providers [8, 17].

Although one could argue that the risks of online self-disclosure are not as severe as the ones from smoking, unwanted incidents such as cyberbullyng, unjustified discrimination and reputation damage should not be neglected or disregarded. Moreover, this urges developers of media technologies and public policy makers to cooperate on behalf of the users' privacy rights. Scholars have suggested that users who are more aware of the consequences of online self-disclosure acts are less likely to share private information in SNSs and more likely to protect their privacy. Likewise, they suggest that in order to inspire risk reduction behaviours, risks should be perceived as controllable by the users [26]. Therefore, risk communication and management is a strategy that deserves to be explored not only for the development of preventative technologies but also for shaping public policies that promote privacy awareness in SNSs.

The interface shown in Fig. 6 is an example of how privacy heuristics can be used for privacy support in SNSs. In this mock-up, an awareness mechanism leverages the knowledge inside the PHDB to inform the user about the risks associated with his/her post. Furthermore, it recommends the right audience for it using the DAT of the corresponding heuristic. In order to follow the *adaptive* principle introduced in Section III-A, the frequency and intensity of these interventions should be regulated. One alternative is to use the frequency level that the user has assigned to the risk to adapt these values (e.g. less and softer warnings for minor risks, and stronger frequent ones for risks with higher severity). Other adaptation variables like the user's privacy concerns [27] and the number of times a user accepts or rejects a suggestion could also be considered for this purpose.

Fig. 6. Envisioned application of heuristics

VI. Conclusion and Future Work

Privacy heuristics are a promising approach for the development of adaptive, visceral, and supportive preventative technologies. They allow users to foresee the potential risks of revealing private and sensitive information in SNSs, and consequently to take more informed privacy decisions. In this work we have introduced DATs for specifying the audience component of privacy heuristics taking into account user-centred privacy preferences. This approach allows to personalize heuristics with minor intervention from the users and recommend access-control decisions more aligned with the users' privacy expectations.

Moving into a direction that goes beyond the theoretical foundations of this approach is a work in progress. During this process we have identified several challenges related to the evaluation of our proposal. The most important one is related to the data required for developing a functional prototype. Basically, one must count with a network of users characterized with social attributes in order to shape privacy heuristics. This has become quite challenging during the last years since SNSs like Facebook have restricted the access to the information inside their social graph through their APIs. Currently, one can access to his/her own friend list trough the Facebook Graph API, but attributes like age, gender, or location that are necessary for the construction of a DAT cannot be retrieved [28].

Due to the technical limitations above mentioned, we have decided to move towards a simulation model of SNSs combining random graphs and *homophily*. Homophily is a basic organization principle in which people with similar characteristics tend to create connections with each other inside a network [29]. It is a process that takes place in the real world and in SNSs which affects not only the type and strength of online relationships but also the users' privacy choices. Our future work consists of a simulation model of privacy choices based on homophily principles. Through this model, we expect to evaluate not only our approach but also the influence of homophily in the performance of other models for audience prediction.

Trust is another aspect that will be part of our future research. Recent news on a social media analytics firm who collected data from 50 million Facebook's users in a dubious, possible illegal ways, has damaged the trust and reputation of SNSs in a severe way [30]. Moreover, such events have also damaged the trust on recommender systems within the social media ecosystem which are often seen as a privacy threats and not as instruments for improving users' online experience. The approach discussed in this paper is an example of how users' data can be used on behalf of their privacy integrity, and not for generating a profit out of it. However, this can awake several privacy concerns regarding possible data misprocessing and leakage in SNSs. Therefore, users' trust on access-control recommendations together with ethical considerations on data collection and processing are aspects to be analysed and discussed in further publications.

VII. Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

We would like to thank Prof. Alfred Kobsa from the Donald Bren School of Information and Computer Sciences who provided insight and expertise that greatly assisted this research.

References

[1] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor, "I regretted the minute I pressed share: A Qualitative Study of Regrets on Facebook," in *Proceedings of the Seventh Symposium on Usable Privacy and Security, SOUPS 2011*. ACM, 2011.

[2] H. R. Lipford, A. Besmer, and J. Watson, "Understanding Privacy Settings in Facebook with an Audience View," in *Proceedings of the 1st Conference on Usability, Psychology, and Security*. USENIX Association, 2008, p. 2.

[3] K. Strater and H. R. Lipford, "Strategies and Struggles with Privacy in an Online Social Networking Community," in *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 1*. British Computer Society, 2008, pp. 111–119.

[4] J. H. Ziegeldorf, M. Henze, R. Hummen, and K. Wehrle, "Comparison-based Privacy: Nudging Privacy in Social Media (position paper)," in *International Workshop on Data Privacy Management*. Springer, 2015, pp. 226–234.

[5] L. Fang and K. LeFevre, "Privacy Wizards for Social Networking Sites," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 351–360.

[6] N. E. Díaz Ferreyra, J. Schäwel, M. Heisel, and C. Meske, "Addressing Self-disclosure in Social Media: An Instructional Awareness Approach," in *Proceedings of the 2nd ACS/IEEE International Workshop on Online Social Networks Technologies (OSNT)*. ACS/IEEE, December 2016.

[7] N. E. Díaz Ferreyra, R. Meis, and M. Heisel, "Should User-generated Content be a Matter of Privacy Awareness? A position paper," in *Proceedings of the 9th International Conference On Knowledge Management and Information Sharing (KMIS 2017)*, November 2017.

[8] S. Samat and A. Acquisti, "Format vs. Content: The Impact of Risk and Presentation on Disclosure Decisions," in *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. Santa Clara, CA: USENIX Association, 2017, pp. 377–384. [Online]. Available: https://www.usenix.org/conference/soups2017/technical-sessions/presentation/samat-disclosure

[9] H. K. Kim, *Risk Communication*. Cham: Springer International Publishing, 2017, pp. 125–149. [Online]. Available: https://doi.org/10.1007/978-3-319-50530-5_7

[10] G. F. Loewenstein, E. U. Weber, C. K. Hsee, and N. Welch, "Risk as Feelings," *Psychological Bulletin*, vol. 127, no. 2, p. 267, 2001.

[11] N. E. Díaz Ferreyra, R. Meis, and M. Heisel, "Online Self-disclosure: From Users' Regrets to Instructional Awareness," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, September 2017, pp. 83–102.

[12] A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L. F. Cranor, S. Komanduri, P. G. Leon, N. Sadeh, F. Schaub, M. Sleeper *et al.*, "Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online," *ACM Computing Surveys (CSUR)*, vol. 50, no. 3, p. 44, 2017.

[13] Y. Wang, P. G. Leon, K. Scott, X. Chen, A. Acquisti, and L. F. Cranor, "Privacy Nudges for Social Media: An Exploratory Facebook Study," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 763–770.

[14] J. Schäwel, "Paving the Way for Technical Privacy Support: A Qualitative Study on Users' Intentions to Engage in Privacy Protection," in *The 67th Annual Conference of the International Communication Association*, 2017.

[15] G. Misra and J. M. Such, "REACT: REcommending Access Control Decisions To Social Media Users," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ser. ASONAM '17. New York, NY, USA: ACM, 2017, pp. 421–426.

[16] G. Misra, J. M. Such, and H. Balogun, "Non-Sharing Communities? An Empirical Study of Community Detection for Access Control Decisions," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 49–56.

[17] L. Stark, "The Emotional Context of Information Privacy," *The Information Society*, vol. 32, no. 1, pp. 14–27, January 2016.

[18] N. E. Díaz Ferreyra, R. Meis, and M. Heisel, "Towards an ILP Approach for Learning Privacy Heuristics From Users' Regrets," in *Proceedings of the 4th European Network Intelligence Conference (ENIC)*, September 2017.

[19] S. Muggleton and L. De Raedt, "Inductive Logic Programming: Theory and Methods," *The Journal of Logic Programming*, vol. 19, pp. 629–679, 1994.

[20] C. Kingsford and S. L. Salzberg, "What are decision trees?" *Nature biotechnology*, vol. 26, no. 9, pp. 1011–1013, 2008.

[21] T. M. Mitchell, *Machine Learning*, ser. McGraw-Hill Series in Computer Science. McGraw-Hill Education, March 1997.

[22] T. B. N. Hoang, V. Moriceau, and J. Mothe, "Predicting Locations in Tweets," in *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CLICLing 2017)*, April 2017.

[23] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of database systems*. Springer, 2009, pp. 532–538.

[24] L. Rokach and O. Maimon, *Data Mining With Decision Trees: Theory and Applications*, 2nd ed. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2014.

[25] H. Hiilamo, E. Crosbie, and S. A. Glantz, "The Evolution of Health Warning Labels on Cigarette Packs: The Role of Precedents, and Tobacco Industry Strategies to Block Diffusion," *Tobacco control*, vol. 23, no. 1, pp. e2–e2, 2014.

[26] E. Christofides, A. Muise, and S. Desmarais, "Risky Disclosures on Facebook: The Effect of Having a Bad Experience on Online Behavior," *Journal of adolescent research*, vol. 27, no. 6, pp. 714–731, 2012.

[27] N. K. Malhotra, S. S. Kim, and J. Agarwal, "Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model," in *Information Systems Research*. Informs, December 2004, vol. 15, no. 4, pp. 336–355.

[28] Facebook, "Facebook Graph API Reference," https://developers.facebook.com/docs/graph-api/reference, (online) Last access on February 18, 2018.

[29] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[30] economist.com, "Epic fail: Facebook faces a reputational meltdown," Online, March 22nd 2018.

DECLARATION OF INDIVIDUAL CONTRIBUTIONS

Publication title: Access-Control Prediction in Social Network Sites: Examining the Role of Homophily

Reference item: N. E. Díaz Ferreyra, T. Hecking, H. U. Hoppe, and M. Heisel, " Access-Control Prediction in Social Network Sites: Examining the Role of Homophily," in *10th International Conference on Social Informatics (SocInfo)*, September 2018. Springer LNCS, pp. 61-74. doi: 10.1007/978-3-030-01159-8_6

Author	Contribution	%
Nicolás E. Díaz Ferreyra	- Conceptualisation of the approach. - Planification of the work. - Draft of the manuscript. - Implementation of the approach.	75%
Tobias Hecking	- Discussion of the approach. - Supervision and advice.	10%
H. Ulrich Hoppe	- Discussion of the approach. - Supervision and advice.	10%
Maritta Heisel	- Supervision and advice.	5%

Nicolás E. Díaz Ferreyra

Tobias Hecking

H. Ulrich Hoppe

Maritta Heisel

Access-control Prediction in Social Network Sites: Examining the Role of Homophily

Nicolás E. Díaz Ferreyra [1], Tobias Hecking[2], H. Ulrich Hoppe[3], and Maritta Heisel[4]

University of Duisburg Essen, Germany
RTG User-Centred Social Media
https://www.ucsm.info/
{nicolas.diaz-ferreyra[1], maritta.heisel[4]}@uni-due.de
{hecking[2], hoppe[2]}@collide.info

Abstract. Often, users of Social Network Sites (SNSs) like Facebook or Twitter have issues when controlling the access to their content. Access-control predictive models are used to recommend access-control configurations which are aligned with the users' individual privacy preferences. One basic strategy for the prediction of access-control configurations is to generate access-control lists out of the emerging communities inside the user's ego-network. That is, in a *community-based* fashion. Homophily, which is the tendency of individuals to bond with others who hold similar characteristics, can influence the network structure of SNSs and bias the users' privacy preferences. Consequently, it can also impact the quality of the configurations generated by access-control predictive models that follow a community-based approach. In this work, we use a simulation model to evaluate the effect of homophily when predicting access-control lists in SNSs. We generate networks with different levels of homophily and analyse thereby its impact on access-control recommendations.

Keywords: homophily · preferential attachment · adaptive privacy · access-control prediction · social network sites

1 Introduction

Users of Social Network Sites (SNSs) like Facebook or Twitter interact with a vast network of people who often represent various facets of their life. Like in the real world, similarities in gender, age, race, nationality or education level gather together people in online communities. This basic organization principle, in which people with similar characteristics tend to create connections with each other inside a network, is called *homophily* or *assortative mixing* [13]. In terms of structural properties, homophily translates into attribute similarity (i.e. "closeness" in terms of profile attributes) between the actors inside a network [16, 18]. Consequently, it affects the type and strength of online relationships, and the formation of communities or clusters in a SNS.

The type and strength of online social relationships are considered important factors which influence users' access-control decisions inside SNSs [15]. For

instance, a user Alice can choose to exclude her work colleagues from a negative post she writes about her workplace in order to avoid future problems with her employer. As one can see in this example, feature similarity (in this case "workplace") is used by the user as an access-control *rule of thumb* in order to protect her privacy. Although this is a good and common privacy strategy, creating and maintaining access-control lists or circles can generate a high cognitive burden on the users. In consequence, users do not employ these privacy-preserving features, leaving their private information accessible to unintended audiences.

Privacy scholars have engineered different Access-control Predictive Models (ACPMs) in order to relieve the users from the burden of creating and maintaining lists of information recipients [14, 7, 6]. These models generate and recommend Access-control Lists (ACLs) based on the users' previous *privacy decisions* (i.e. with whom they have shared private information in the past), or their *privacy preferences* (i.e. asking them to assign *allow/deny* labels to a representative sample of friends). In other words, they predict *black lists* of information recipients using the user's privacy preferences or past decisions as predictor *variables*, and particular *instances* of these predictors as training examples. One basic strategy is to map these examples to emerging communities inside the user's *ego-network* (i.e. the network of connections between his/her friends) [15]. For instance, if a user Alice has excluded her friends Bob, John and Bill from the audience of a particular post, then it is to expect that she will exclude them again in the future together with other friends with similar characteristics. Let us assume that Alice's ego network can be clustered into three communities $C1$, $C2$ and $C3$, where Bob, John and Bill are grouped together in $C1$ (as illustrated in Fig. 1). Since community membership often indicates similarity between people, then one could expect that the rest of friends inside community $C1$ are also ought to be excluded by Alice from the audience of her future posts. Hence, an ACL consisting of all members of $C1$ is generated and recommended to Alice.

Community-based ACPMs like the one just mentioned follow a *guilt-by-association* approach [4, 15]. This is, all members inside a community are classified as untrusted recipients of a particular post because one fraction of its members were previously classified as untrusted (e.g. in Fig. 1 Amanda, Marc, Susan, Sarah and Mary are guilty-by-association). Although this approach is a

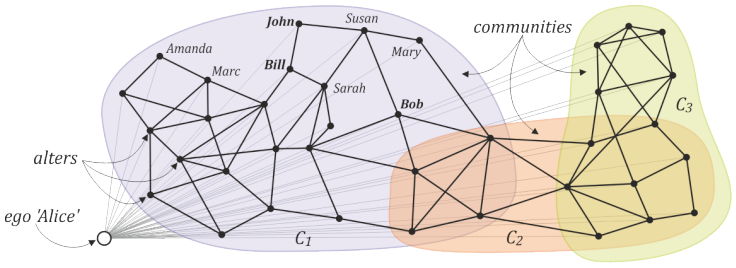


Fig. 1. Guilt-by-association approach for access-control prediction in SNSs.

good starting point for recommending personalized access-control lists, its accuracy depends largely on the semantics of these untrusted communities. For instance, if Alice excludes Bob, John and Bill from a post because they are her work colleagues, then Alice would expect an access-control recommendation containing a high number of work colleagues. However, algorithms used to identify communities inside a network do not take into account attribute similarity between individuals (i.e. homophily) and therefore do not provide cues about the semantic of emerging communities. In other words, for the example of Fig. 1 it is not possible to determine whether $C1$ gathers people working in the same company, people living in the same city, or people who share the same music taste. Therefore, it is hard to guarantee that the recommended access-control list will contain a large number of Alice's work colleagues. Consequently, there is a probability for Amanda, Marc, Susan, Sarah and Mary to be unfairly guilty-by-association.

As it is shown in the previous example, homophily between individuals can influence the structural properties of ego-networks and impact the performance of access-control predictions. In this work we examine the role of homophily when predicting ACLs in SNSs though a simulation model. Using this model, we generate (i) ego-networks out of different homophily scenarios, (ii) user-centred privacy preferences, and (iii) ACLs following the guilt-by-association approach. We show that community-based ACPMs can lead to unfair predictions under particular homophily scenarios, and that ACLs require validation from the user to ensure that they are aligned with his/her privacy preferences.

The rest of this paper is organized as follows. In the next section we briefly outline the simulation methodology used to analyse the impact of homophily when predicting ACLs in SNSs. The simulation of ego networks, which is an important part of our approach, is fully explained in the Appendix. In section 3, we use our methodology to generate ACLs under different homophily scenarios and evaluate their performance against user-centred privacy preferences. For this, we propose a fairness metric and compare the results of each simulation execution. In section 4 we analyse the strengths and limitations of our method. Finally, in section 5 we conclude and discuss future work.

2 Methodology

As it is shown in Fig. 2, an ACPM requires (i) the user's ego-network and (ii) his/her privacy preferences in order to generate a personalized ACL. To simulate the user's ego-network, we developed a preferential-attachment model for generating scale-free networks. In this model, the ego and its alters are characterized with the attributes *gender*, *workplace* and *location* where gender can take the values *male* or *female*, workplace the values *Starbucks*, *Google* or *Ikea*, and location the values *Leeds* or *York*. The preferential attachment rule takes into account an openness matrix $\boldsymbol{\Lambda}$ to compute the linking probability between two nodes. The values inside $\boldsymbol{\Lambda}$ can range from 0 to 1 and describe how strong/weak attribute similarity is for this linking process (e.g. a value Λ_{York}^{Leeds} closer to zero

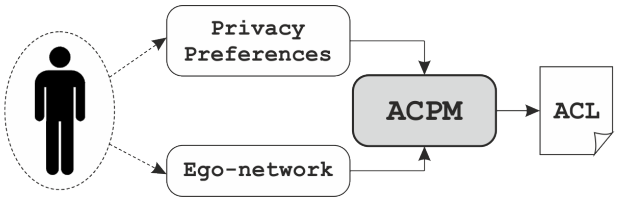


Fig. 2. Simulation Pipeline.

describes a setting in which users located in *York* are less likely to connect with users living in *Leeds*). In this work, we have analysed the role of homophily in community-based ACPMs through different configurations of $\boldsymbol{\Lambda}$ which represent different homophily scenarios. A full description of this model can be found in the Appendix.

In addition to an ego-network, a community-based ACPM requires the user's privacy preferences to generate a personalized ACL. In the practice, such preferences are elicited by asking the user to assign *allow/deny* labels to a reduced sample of friends. That is, the user is asked to provide examples of contacts that should be excluded/included from the audience of a particular piece of private information (e.g. a post or profile attributes). Basically, this kind of access-control decisions are influenced by two factors (i) the homophily degree between the user and his/her contacts, and (ii) the type of information to be disclosed. For instance, a user who posts a negative comment about his/her workplace is likely to assign *deny* labels to a sample of work colleagues inside his/her network. Then, using the emerging communities inside the user's network, the ACPM takes this set of examples and generates a larger customized ACL.

In order to simulate the user's privacy preferences, we must introduce a *selection criterion* of nodes inside the simulated ego-network. Given a *self-disclosure scenario* (i.e. a *user profile* and a piece of *private information* to be disclosed), such criterion consists of selecting nodes that hold a *critical attribute*. For instance, let us assume that our user is *male*, works at *Ikea* and lives in *York*, and the information to be disclosed is a post containing a negative comment about the user's workplace. If this post is seen by a colleague from *Ikea* the user might suffer an unwanted incident such a reputation damage, wake-up call from a superior or even job loss. Therefore, the attribute *Ikea* is *critical* for the user's privacy under this scenario. We can simulate the privacy preferences for this scenario by selecting n nodes from the simulated ego-network that hold the critical attribute *Ikea*. In order to choose the most influential nodes among these ones, we will select the n nodes with the highest degree.

3 Simulation Execution

In order to analyse the role of homophily in community-based ACPMs we put our methodology into practice. For this, we simulate ego-networks with different

homophily configurations and use them to generate ACLs. Thereafter, we analyse the accuracy of these ACLs through the privacy preferences derived from a particular self-disclosure scenario.

3.1 Homophily Scenarios

As mentioned previously, each homophily scenario is described using a $\boldsymbol{\Lambda}$ matrix (refer to Appendix B for details). In our case, we have chosen similarity in *location* and *workplace* as the homophily aspects to control in our simulations. Therefore, the $\boldsymbol{\Lambda}$ of these scenarios differ only in the values assigned to Λ_{York}^{Leeds}, $\Lambda_{Google}^{Starbucks}$, $\Lambda_{Starbucks}^{Ikea}$ and Λ_{Google}^{Ikea} while the rest of the group-openness factors are set to one. The homophily scenarios used to execute our model were the following:

- $S_1 : \{\Lambda_{Google}^{Starbucks} = \Lambda_{Starbucks}^{Ikea} = \Lambda_{Google}^{Ikea} = 0.7\}$
- $S_2 : \{\Lambda_{Google}^{Starbucks} = \Lambda_{Starbucks}^{Ikea} = \Lambda_{Google}^{Ikea} = 0.3\}$
- $S_3 : \{\Lambda_{Google}^{Starbucks} = \Lambda_{Starbucks}^{Ikea} = \Lambda_{Google}^{Ikea} = 0.01\}$
- $S_4 : \{\Lambda_{York}^{Leeds} = 0.7\}$
- $S_5 : \{\Lambda_{York}^{Leeds} = 0.3\}$
- $S_6 : \{\Lambda_{York}^{Leeds} = 0.01\}$

As it can be observed, we have varied the openness values between the groups corresponding to *workplace* in scenarios S_1, S_2 and S_3. This set of scenarios represent different levels of users' openness/closeness when creating ties with individuals from other workplaces. Likewise, for scenarios S_4, S_5 and S_6 we have varied the openness values between the groups corresponding to *location*. In this case, these scenarios describe different levels of openness/closeness when users bind with individuals from another location.

3.2 Execution and Analysis

We have implemented our simulation model using iGraph [5], a library for network analysis and visualization for R[1]. The model was initialized and executed according to the set-up parameters described in section B.2 of the Appendix and the homophily scenarios introduced previously. We have simulated 6 topologies of ego-networks corresponding to the scenarios S_1, S_2, S_3, S_4, S_5, and S_6. Additionally, we have introduced a control scenario S_0 in which homophily does not influence the preferential attachment rule (i.e $\boldsymbol{\Lambda} = 1$, an all-ones matrix). This scenario was also simulated, giving a total of 7 simulated ego-networks.

As we described in section 2, one must define a *self-disclosure scenario* in order simulate the user's privacy preferences. For our simulations, we have proposed a scenario in which a user (i.e. the ego of the simulated ego-network) who is *male*, works in *Ikea* and lives in *York*, posts a *negative comment* about his employer on a SNS. Based on this scenario, we have simulated the privacy

[1] The scripts can be found in the following repository: https://bit.ly/2Nth7HE

preferences of this user by selecting 10 nodes from each simulated network that hold the critical attribute *Ikea*. This selection was made so that the nodes with the highest degree were chosen. The column *Preferences* of Table 1 shows the privacy preferences generated for each simulated scenario. We can observe for instance, that for scenario S_0 the 10 nodes with highest degree that hold the attribute $Ikea$ are P_0: {4,6,2,3,28,20,44,65,88,17}. Hence, these nodes are defined as the user's privacy preferences in S_0.

Scn.	Preferences	Method	N° Communities	Size of best-fit community	Nodes with critical attribute	Fairness
S_0	P_0:{4,6,2,3,28,20, 44,65,88,17}	MC	15	51	19	37.25%
		LE	11	225	82	36.44%
S_1	P_1:{2,6,4,30,32, 15,75,35,40,25}	MC	15	53	16	31.19%
		LE	28	49	24	48.98%
S_2	P_2:{1,4,11,3,75, 27,13,14,81,10}	MC	15	67	34	50.75%
		LE	16	40	30	75.00%
S_3	P_3:{1,6,16,17,28, 2,54,72,120,78}	MC	6	141	140	99.29%
		LE	6	172	168	97.67%
S_4	P_4:{7,3,8,19,24, 2,14,5,39,55}	MC	15	41	18	43.90%
		LE	18	62	20	32.26%
S_5	P_5:{2,8,5,30,9, 82,18,39,50,23}	MC	15	71	26	36.62%
		LE	20	96	35	36.46%
S_6	P_6:{4,1,28,7,29, 8,18,49,41,5}	MC	9	117	35	29.91%
		LE	9	145	57	39.31%

Table 1. Simulation results for scenarios S_0, S_1, S_2, S_3, S_4, S_5 and S_6.

In order to predict the corresponding ACLs, preferences (i.e. the selected nodes) are mapped to the emerging communities inside the ego-network of each scenario. For each scenario, the community that best-fits the preferences (i.e. the one that covers the majority of the nodes) is then chosen as ACL. To identify communities inside networks we have applied two different community-detection algorithms: Leading Eigenvector (LE) [17] and Multilevel Community (MC) [3]. Both methods generate a hierarchy of nested communities in which nodes belong to at most one community. In the case of LE, this process is done following a top-down approach (i.e. starting with the entire network and iteratively splitting it into partitions), whereas in the case of MC, this is done in a bottom-up fashion (i.e. starting with single nodes and iteratively merging them into communities). Both algorithms are part of the iGraph package.

As described in section 2, the community that best-fits the user's privacy preferences becomes the personalized ACL. In order to determine if such ACL is indeed aligned with the user's privacy preferences, we have defined a *fairness* metric. In this case, fairness is computed as the percentage of nodes that hold the critical attribute inside the predicted ACL. For instance, when using MC the community that best-fits P_0 has 19 out of 51 nodes with the attribute $Ikea$. Therefore, the fairness of the corresponding ACL is $19/51 * 100 = 37.25\%$. Low

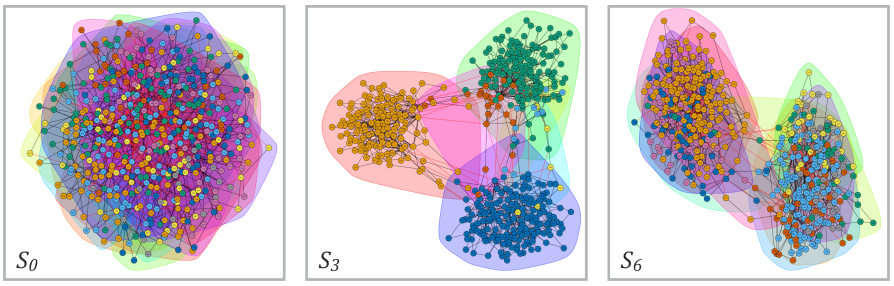


Fig. 3. Emerging communities for scenarios S_0, S_3 and S_6 using MC.

levels of fairness like this one indicate that the predicted ACL is not reflecting the privacy expectations of the user. That is, instead of generating a list of contacts mainly composed by work colleagues from $Ikea$, the ACPM generates a list mostly composed of contacts who do not work there. Hence, the ACL will not cover the privacy expectations and needs of the user, and most of the contacts inside the ACL will be unfairly guilty-by association.

Table 1 shows that the highest and lowest levels of fairness correspond to scenarios S_3 and S_6, respectively. The emerging communities for these configurations together with the ones of S_0 are shown in Fig. 3. As it can be observed, the plot corresponding to S_3 has three highly dense areas and the one from S_6 has two. This is due to the $\boldsymbol{\Lambda}$ assigned to each scenario: in S_3 nodes with *workplace* similarity were more likely to connect to each other, whereas in S_6 this was the case for nodes with the same *location*. Consequently, in S_3 nodes are condensed in areas corresponding to the values $Ikea$, $Starbucks$ and $Google$ whereas nodes in S_6 are condensed in areas corresponding to $Leeds$ and $York$. Therefore, since emerging communities in S_3 reflect *workplace* similarity, the fairness of its corresponding ACL is much higher than in S_6 where communities reflect *location* similarity. Consequently, in S_3 the effect of homophily contributes positively to the fairness of the predicted ACL, whereas in S_6 such effect has a negative impact.

4 Discussion

From the execution of our model, we can conclude that community-based ACPMs can lead to unfair predictions under certain homophily scenarios. Basically, we have observed that this happens when the semantic of the best-fit community is not aligned with the user's privacy preferences. However, one should take into account that in a controlled simulation environment (like the one used for this study) assumptions and parameters can influence the results of our experiments. One of the assumptions that should be closely analysed is the distribution of attributes proposed in section B.1 of the Appendix. For our experiments, we have proposed three attributes and distributed their values uniformly. However, this

is not always the case and often some values tend to prevail over others. For instance, Volkovich et al. [19] suggest that online ties preferentially connect closer people. For the self-disclosure scenario that we have proposed, this suggest that nodes with $location = York$ should prevail over nodes with $location = Leeds$. Consequently, the distribution of attributes should be characterized in a way that resembles a more realistic set-up.

Another aspect to be considered is related to the homophily scenarios proposed to generate the ego-networks. We have considered basically two homophily factors in our simulations: (i) similarity in workplace and (ii) similarity in location. The scenarios proposed in section 3.1 consider homophily as an *attribute-specific* phenomenon and not as a *cross-attribute* phenomenon. That is, group-openness factors were specified only among the groups corresponding to a particular attribute and not between groups from different attributes. For instance, scenarios S_1, S_2 and S_3 define values (different to one) for Λ_{York}^{Leeds}, $\Lambda_{Google}^{Starbucks}$ and $\Lambda_{Starbucks}^{Ikea}$, but do not define values for Λ_{York}^{Ikea} or $\Lambda_{Leeds}^{Starbucks}$. Therefore, the influence of homophily across the attributes *workplace* and *location* is ignored in our simulations. It is necessary to take a closer look at the predominant homophily scenarios in SNSs and incorporate this information to the model. Empirical evidence can be a vehicle to cope with this task.

5 Conclusions and Future Work

The limitations of community-based ACPMs that we have exposed in this work suggest that users should take a more active role in the automatic generation of ACLs. Basically, ACPMs should include a validation stage in which the user can provide feedback on the composition of the resulting ACL. Strategies to incorporate users into the information flow of recommender systems is one of the main challenges of human-computer interaction [8]. Our future work will focus on exploring different alternatives to include the user's feedback into ACPMs for SNSs. Likewise, we will investigate the predominant homophily processes that take place on SNSs in order to improve the efficiency of our model.

The community-detection algorithms used for the simulation are also a matter of future research. Basically, the two approaches used in this work identify non-overlapping communities based solely on the network structure. Approaches like the one of Leskovec et al. [11] generate overlapping communities combining network structure with attribute similarity. We believe this can help to discover communities that adjust better to the user's privacy preferences and improve thereby the fairness of the generated ACLs. This aspect will be analysed and discussed in future publications.

Acknowledgments. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

Appendix

In this Appendix we introduce the theoretical foundations used for the definition and implementation of our simulation model of ego-networks.

A Network Evolution with Homophily

Up to now, scholars have proposed several evolution models for constructing scale-free networks. Among them, the *prefferential attachment* mechanism introduced by Barabassi and Albert [2] is one of the most prominent ones. However, this model does not consider attribute similarity when computing the linking probability between two nodes. Following, we introduce an approach for the simulation of ego-networks that takes homophily explicitly into account.

A.1 Group-openness

Many empirical and theoretical studies have shown that people prefer to link to those people with whom they share certain characteristics [13]. Moreover, these studies have also shown that homophily can lead to the emergence of clusters inside a social network in which similar people are linked more densely with each other [9]. In order to study the role of homophily in the evolution of scale-free networks, Kim et al. [10] introduced *group-openness* characteristics to the preferential attachment mechanism of Barabasi and Albert [1]. In this approach, nodes which share a particular characteristic s belong to group s. Consequently, the group-openness factor Λ_s^t between two groups s and t is defined as:

$$\Lambda_s^t = \begin{cases} \Lambda, & \text{if } s \neq t \\ 1, & \text{if } s = t \end{cases} \qquad 0 \leq \Lambda \leq 1 \tag{1}$$

where the homophily index Λ is a real number between 0 and 1 [9]. If $\Lambda = 0$, nodes in group s do not link with nodes in group t ($t \neq s$) but link only with those nodes in the same group. This state describes completely *closed* groups, in which members prefer to link only with those who hold their same characteristics. Conversely, if $\Lambda = 1$ homophily does not affect the linkage between nodes, independently of whether they belong to the same group or not. This state describes completely *open* groups that show neutrality when linking to others [10].

A.2 Preferential Attachment with Homophily

The preferential attachment mechanism introduced by Barabasi and Albert [1] describes the process by which new nodes prefer to link to the more connected nodes in a network (i.e. the hubs). Hence, the probability Π_i that a new node connects to node i is proportional to the degree k_i of node i:

$$\Pi_i = \frac{k_i}{\sum_j k_j}.m \qquad \begin{array}{l} m\text{=number of new links} \\ k_i\text{= degree of node } i \end{array} \tag{2}$$

Using the group-openness mechanism defined in Eq. 1, Kim et al. [10] introduced homophily to the preferential attachment model of Eq. 2. As result, the probability Π_i^{pq} that a new node of group q is linked to node i of group p is defined as:

$$\Pi_i^{pq} = \frac{k_i^p.\Lambda_p^q}{\sum_j k_j^\mu.\Lambda_\mu^q} \qquad \begin{array}{l} k_i^p\text{= degree of node } i \text{ from group } p \\ \Lambda_p^q\text{= opennes of group } p \text{ with group } q \end{array} \tag{3}$$

where k_i^p is the degree of node i of group p, and Λ_p^q represents the homophily between the group of the new node q and the group of node i. As one can observe, in Eq. 2 the probability that a new node connects to an existing node i is normalized by the sum of degrees of all existing nodes in the network. This is also the case for Eq. 3, only that this time the group-openness factor of each node is considered for the normalization. In other words, if all groups are completely open (i.e. $\Lambda = 1$), Eq. 3 is identical to Eq. 2. On the other hand, if the groups are completely open (i.e. $\Lambda = 0$), Eq. 3 is reduced to Eq. 2 for each particular group [10].

In order to explain the evolution of nodes who are active inside a network for a long period of time, Kim et al. [9] introduced an additional rule which describes the creation of links between existing nodes. This is, the probability Π_{ij}^{pq} that node i of group p links to node j of group q is defined as:

$$\Pi_{ij}^{pq} = \frac{k_i^p.k_j^q.\Lambda_p^q}{\sum_l \sum_{m>l} k_l^\mu.k_m^\nu.\Lambda_\mu^\nu} \tag{4}$$

where k_i^p and k_j^q are the degrees of node i and of node j respectively [10]. Nodes i and j belong to groups p and q respectively, and Λ_p^q is the group-openness between these two groups. Like in Eqs. 2 and 3, Eq. 4 is normalized by the sum of all possible combinations of links between existing nodes in the network [10]. In this case, k_l^μ and k_m^ν are the degrees of nodes l and m which belong to groups μ and ν, respectively. Likewise, Λ_μ^ν refers to the group-openness between groups μ and ν [10].

B Simulation Model

The model introduced in Section A.2 of this Appendix generates a network in which nodes link with each other according to attribute similarity. Therefore, it assumes that the values of these attributes have been assigned to the nodes prior to the attachment phase. Following, we define the attributes used to characterise the nodes of our simulated networks and the distribution of their respective values. Likewise, we define the parameters used to set-up the simulation.

B.1 Node-attributed Ego-Networks

In our model, the ego and its alters are characterized with the attributes *gender*, *workplace* and *location* where gender can take the values *male* or *female*, workplace the values *Starbucks*, *Google* or *Ikea*, and location the values *Leeds* or *York*. These attributes and their respective values are conditionally distributed following the probability tree of Fig. 4. According to this distribution, a node in the network is generated with 50% chance of being female and 50% chance of being male (i.e. $P(male) = P(female) = 0.5$). Then, the values for workplace are assigned with 33% chance according to the gender value of the node. For instance, if $gender = female$, then $P(Starbucks|female) = P(Google|female) = P(Ikea|female) = 0.33$. Likewise, the values for location are assigned with 50% chance given the gender and workplace values of the node. This means that in the case of a node whose gender and location attributes are *female* and *Ikea*, then $P(York|female \cap Ikea) = P(Leeds|female \cap Ikea) = 0.5$.

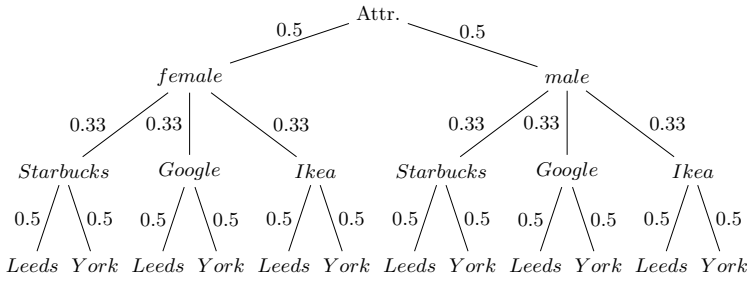


Fig. 4. Attributes probability distribution.

Each attribute value represents a group. Therefore, our model consists of 7 groups (i.e. *male*, *female*, *Starbucks*, *Google*, *Ikea*, *Leeds* and *York*) together with the corresponding group-openness factors between them. If we consider group-openness a symmetric relation between two groups s and t, then $\Lambda_s^t = \Lambda_t^s$. This means that for 7 groups one must define $C_{7,2} = \frac{7!}{2!(7-2)!} = 21$ different group-openness factors. This information can be expressed through a group-openness matrix $\boldsymbol{\Lambda}_{7x7}$ in which each cell represents a factor Λ_s^t as shown in Fig.5. As one can observe, this matrix is symmetric and contains ones on its main diagonal. This is because $\Lambda_s^t = \Lambda_t^s$ and, according to Eq. 1, the group-openness factor Λ_s^t is 1 when $s = t$.

Nodes are described in terms of one value per attribute and, consequently, belong to more than one group at the same time (e.g. a node whose gender is *female*, works in *Ikea* and lives in *York*, belongs to the groups *female*, *Ikea* and *York* respectively). Therefore, the *total* homophily factor between two nodes i and j depends on more than one group-openness factor. In other words, one should compute the homophily between i and j considering the group-openness

$$\boldsymbol{\Lambda} = \begin{bmatrix} \Lambda_{male}^{male} & \Lambda_{male}^{female} & \dots & \Lambda_{male}^{York} \\ \Lambda_{female}^{male} & \Lambda_{female}^{female} & \dots & \Lambda_{female}^{York} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{York}^{male} & \Lambda_{York}^{female} & \dots & \Lambda_{York}^{York} \end{bmatrix} = \begin{bmatrix} 1 & 0.7 & \dots & 0.3 \\ 0.7 & 1 & \dots & 0.5 \\ \vdots & \vdots & \ddots & \vdots \\ 0.3 & 0.5 & \dots & 1 \end{bmatrix}$$

Fig. 5. Group-openness Matrix.

factors of all possible combinations among the groups to which i and j belong. For instance, if i belongs to the group-set *male*, *Starbucks* and *York*, and j to the group-set *female*, *Google* and *York*, then one should consider Λ_{female}^{male}, Λ_{Google}^{male}, Λ_{York}^{male}, $\Lambda_{female}^{Starbucks}$, $\Lambda_{Google}^{Starbucks}$, $\Lambda_{York}^{Starbucks}$, Λ_{female}^{York}, Λ_{Google}^{York}, and Λ_{York}^{York}. Consequently, the total homophily factor between two group-sets P and Q is defined as:

$$\mathcal{H}_P^Q = \prod_{\substack{p \in P \\ q \in Q}} \boldsymbol{\Lambda}_{p,q} \qquad \begin{array}{l} P\text{= groups to which node } i \text{ belongs} \\ Q\text{= groups to which node } j \text{ belongs} \end{array} \tag{5}$$

where P and Q are the groups to which nodes i and j belong, respectively. According to the definition above, the preferential attachment model described in Eqs. 3 and 4 can be re-defined. That is, the probability Π_i^{PQ} that a new node of group-set Q is linked to node i of group-set P is defined as:

$$\Pi_i^{PQ} = \frac{k_i^P . \mathcal{H}_P^Q}{\sum_j k_j^M . \mathcal{H}_M^Q} \qquad \begin{array}{l} k_i^P\text{= degree of node } i \text{ from group-set } P \\ \mathcal{H}_P^Q\text{= total homophily factor between} \\ \text{group-set } P \text{ and group-set } Q \end{array} \tag{6}$$

where k_i^P is the degree of node i from group-set P, and $\mathcal{H}_P^Q$ represents the total homophily factor between the group-set of the new node Q and the group-set of node i. Likewise, the probability Π_{ij}^{PQ} that node i of group-set P links to node j of group-set Q is defined as:

$$\Pi_{ij}^{PQ} = \frac{k_i^P . k_j^Q . \mathcal{H}_P^Q}{\sum_l \sum_{m>l} k_l^M . k_m^N . \mathcal{H}_M^N} \tag{7}$$

where k_i^P and k_j^Q are the degrees of node i and of node j respectively and $\mathcal{H}_P^Q$ is the total homophily factor between group-sets P and Q.

B.2 Simulation Set-up

Our simulation model for ego-networks comprises Eqs. 6 and 7 together with the attribute probability distribution introduced in Apendix B. According to

the Pew Research Center, the average size of an ego-network in Facebook was of 338 friends/nodes in 2014[2] (this number scaled up to 425 in a study focused on adolescents and online privacy in 2013 [12]). Therefore, we will consider an average ego-network consisting of 500 nodes and execute our simulation for $F = 500$ time units. The initial set-up for all simulations consists of an network of two nodes ($N(0) = 2$) and one link ($K = 1$). It is also assumed that only one node enters the network at time t ($b = 2$) and generates only one link ($\beta = 1$). On the other hand, the number of new links between existing nodes at time t is given by $\lfloor N(t) \cdot \alpha \rfloor$ where $0 \leq \alpha < 1$ and $N(t)$ is the number of links at t. In order to preserve the degree distribution in our simulated networks we adopt $\alpha = 0.001$ as suggested by Kim et al. [9].

References

1. Barabási, A.L.: Network Science. Cambridge University Press (2016)
2. Barabási, A.L., Albert, R.: Emergence of Scaling in Random Networks. Science **286**(5439), 509–512 (1999). https://doi.org/10.1126/science.286.5439.509, http://science.sciencemag.org/content/286/5439/509
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10) (2008)
4. Boyd, D., Levy, K., Marwick, A.: The Networked Nature of Algorithmic Discrimination. In: Data and Discrimination: Collected Essays. Open Technology Institute, New America Washington, DC (2014)
5. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal **Complex Systems**, 1695 (2006), http://igraph.org
6. Díaz Ferreyra, N.E., Meis, R., Heisel, M.: At Your Own Risk: Shaping Privacy Heuristics for Online Self-disclosure. In: Proceedings of the 16th Annual Conference on Privacy, Security and Trust (August 2018)
7. Fang, L., LeFevre, K.: Privacy Wizards for Social Networking Sites. In: Proceedings of the 19th International Conference on World Wide Web. pp. 351–360. WWW '10, ACM, New York, NY, USA (2010). https://doi.org/10.1145/1772690.1772727
8. Gil, M., Pelechano, V., Fons, J., Albert, M.: Designing the Human in the Loop of Self-Adaptive Systems. In: International Conference on Ubiquitous Computing and Ambient Intelligence. pp. 437–449. Springer (2016)
9. Kim, K., Altmann, J., Hwang, J.: The Impact of the Subgroup Structure on the Evolution of Networks: An Economic Model of Network Evolution. In: 2010 INFOCOM IEEE Conference on Computer Communications Workshops. pp. 1–9 (March 2010). https://doi.org/10.1109/infocomw.2010.5466705
10. Kim, K., Altmann, J.: Effect of homophily on network formation. Communications in Nonlinear Science and Numerical Simulation **44**, 482–494 (2017)
11. Leskovec, J., McAuley, J.: Learning to Discover Social Circles in Ego Networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 539–547. Curran Associates, Inc. (2012)

[2] http://www.pewresearch.org/fact-tank/2014/02/03/what-people-like-dislike-about-facebook/

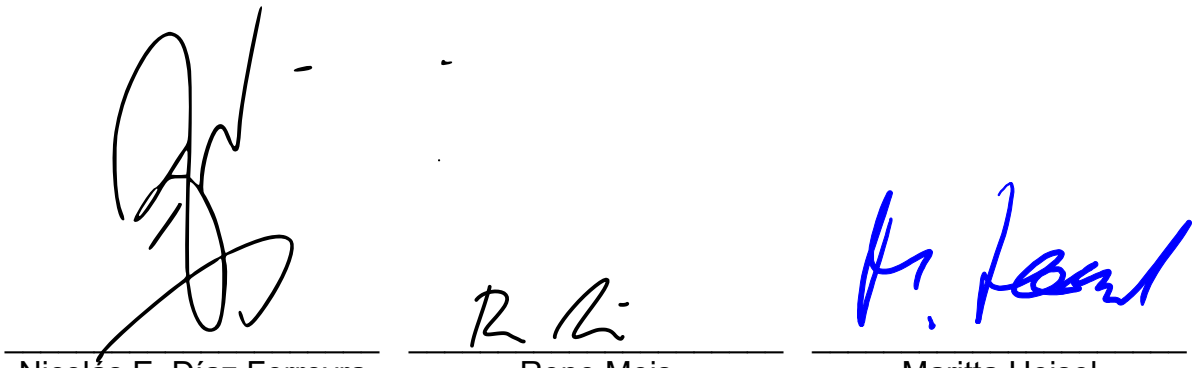
12. Madden, M., Lenhart, A., Cortesi, S., Gasser, U., Duggan, M., Smith, A., Beaton, M.: Teens, Social Media, and Privacy. Pew Research Center **21**, 2–86 (2013)
13. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology **27**(1), 415–444 (2001)
14. Misra, G., Such, J.M.: REACT: REcommending Access Control Decisions To Social Media Users. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. pp. 421–426. ASONAM '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3110025.3110073
15. Misra, G., Such, J.M., Balogun, H.: Non-Sharing Communities? An Empirical Study of Community Detection for Access Control Decisions. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 49–56 (2016)
16. Newman, M.E.J.: Mixing patterns in networks. Physical Review E **67**, 026126 (Feb 2003). https://doi.org/10.1103/PhysRevE.67.026126
17. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Physical Review E **74**, 036104 (Sep 2006). https://doi.org/10.1103/PhysRevE.74.036104
18. Thedchanamoorthy, G., Piraveenan, M., Kasthuriratna, D., Senanayake, U.: Node Assortativity in Complex Networks: An Alternative Approach. Procedia Computer Science **29**, 2449–2461 (2014)
19. Volkovich, Y., Scellato, S., Laniado, D., Mascolo, C., Kaltenbrunner, A.: The Length of Bridge Ties: Structural and Geographic Properties of Online Social Interactions. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (2012)

DECLARATION OF INDIVIDUAL CONTRIBUTIONS

Publication title: Learning from Online Regrets: From Deleted Posts to Risk Awareness in Social Network Sites

Status: Submitted for publication

Author	Contribution	%
Nicolás E. Díaz Ferreyra	- Conceptualisation of the approach. - Planification of the work. - Draft of the manuscript.	90%
Rene Meis	- Supervision and advice.	5%
Maritta Heisel	- Supervision and advice.	5%



Nicolás E. Díaz Ferreyra | Rene Meis | Maritta Heisel