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

N. 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").

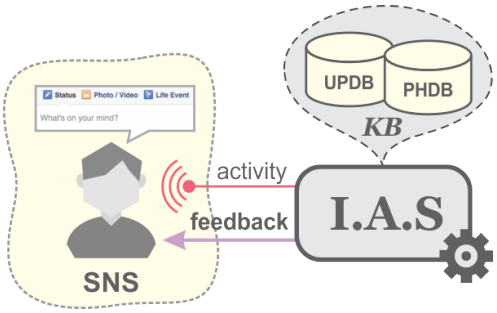

Fig. 1. Instructional Awareness System (IAS)

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.

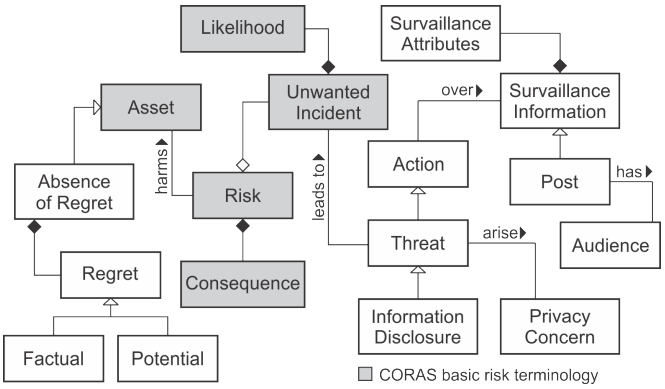

Fig. 2. PHeDer conceptual model

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.

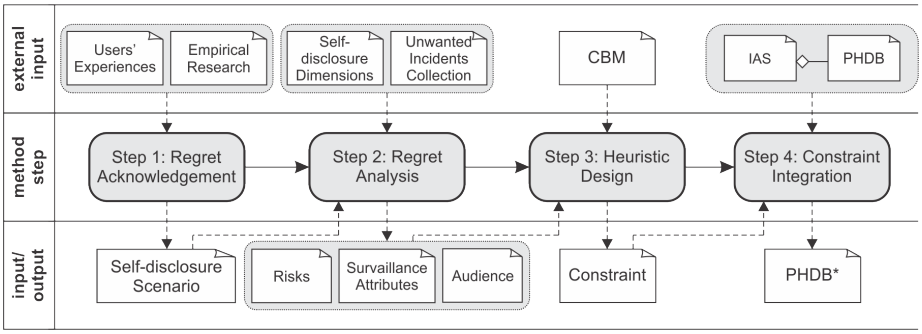

Fig. 3. PHeDer Steps and Artifacts

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)