December 2019

Automatic Generation of Explanations to Prevent Privacy Violations

Ruiz-Dolz RAMON [a,1], Alemany JOSE [a], Heras STELLA [a] and Garcia-Fornes ANA [a]
[a] *Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain*

Abstract. With the massive use of online social environments and technologies, users' concern regarding the privacy of their own data has significantly increased. In this paper, we present a method to automatically generate explanations in a social network domain. This method uses the available data from the network to anticipate any potential privacy violation, automatically generates explanations, and shows them to the user with the purpose of avoiding the detected violation.

Keywords. Argumentation, Explainable AI, Privacy, Social Networks

1. Introduction

With the continuous increase of the population of online social networks [9], users' concern about the privacy of their data has significantly increased. Online social networks provide users with a series of tools for interacting, sharing and publishing information with other users. With these provided tools, to spread information and data is easier than ever, thus it is imperative to make a responsible use of these technologies. When this does not happen, several threats may appear. An important threat to users' privacy may be one's own publications. Although this may not seem very logical, as described in [21], it is very common to find users regretting their own online publications. On the other hand, when involving other users in a publication, since their privacy preferences may be unknown for the author, multi party privacy conflicts [20] are also a very common privacy threat in social networks. Therefore, it is interesting to be able not only to warn users but also to give them explanations of the main reasons of the potential privacy violations detected to minimise their occurrence. Additionally, with the recent appearance of stricter laws in the field of data privacy protection, these privacy violations may also have legal consequences for the privacy violator, and social media users' legal consciousness about privacy is usually very low [16].

This increasing concern about privacy threats in social networks has attracted the interest of researchers to find effective mechanisms to deal with privacy violations. Several privacy management assistance tools have been identified in the literature. In [14,18,19,22], we can observe different approaches to reduce the harm caused by privacy

[1] E-mail: raruidol@dsic.upv.es

violations. In [13], an underlying negotiation protocol to choose the *optimal* privacy policy is presented. However, two important flaws can be identified in all these approaches. First, all of them are focused on minimising conflicts when multiple parties are involved, ignoring the potential self inflicted privacy violations. Second, none of them provide the author with proper explanations of why a specific decision should be made.

As pointed out in [10, 11], argumentation seems to be the most coherent approach to tackle this kind of problems. In [12], a negotiation protocol based on computational argumentation techniques to decide the *optimal* privacy policy is proposed. However, the decision is automatically taken by the system, and no explanation is given to the user.

Computational argumentation has an extensive and successful history of applications in law [5, 6], and it is a suitable a method for enforcing privacy and fairness. However, it has not been until recently that the community has started to investigate the potential applications of argumentation as a tool to provide AI systems with greater explanatory power [17], also in the legal domain [4].

Therefore, with this work we intend to pave the way for the automatic generation of explanations in the privacy management domain. Thus, we propose an argumentation-based approach to automatically generate such explanations and prevent privacy violations in social networks. With our system, users can receive warnings and explanations from the social network site that will improve their awareness on the potential consequences of their publications. The paper is structured as follows, Section 2 contextualises the framework of the research presented in this work. Section 3 presents the method proposed to automatically generate explanations to prevent privacy violations in social networks. In Section 4, a case study is set to illustrate how the proposed method works. Finally, Section 5 summarises the main concepts presented in this work and proposes the main lines of future work.

2. Background

The research carried out in this work is based on two main pillars: PESEDIA, an educational social network which is the application domain of the method proposed in this work, and an argumentation framework for online social networks.

PESEDIA is an online social network for educational and research purposes that includes: (i) the design and development of new metrics to analyse and quantify privacy risks [3], (ii) the application of methods to change users' behaviour regarding their privacy concerns, (iii) the implementation of new features to improve the management of users' content and (iv) the evaluation and testing of new proposals with real users. The underlying implementation of PESEDIA uses Elgg [8], which is an open source engine that is used to build social environments. The environment provided by this engine is similar to other social networks (e.g., Facebook).

In [15], we formally defined an argumentation framework for online social networks and proposed an architecture for an argumentation system capable of handling the whole argumentation process. The proposed architecture is structured in four different modules: (i) the *feature extraction*, (ii) the *argument generation*, (iii) the *solver* and (iv) the *dialogue* modules. The *feature extraction* module is in charge of retrieving all the needed data to both, prevent any potential privacy violation and to determine which computational arguments (i.e., 3-element tuples) will be subsequently generated by the *argument*

generation module. This module must be able to generate four different types of arguments: Privacy, Trust, Risk and Content arguments. Once all the computational arguments are generated, the *solver* module must compute the set of acceptable arguments in favour or against making a publication. When the set of acceptable arguments is against making a publication (meaning that a potential privacy violation has been detected), the *dialogue* module is in charge of trying to persuade the author to modify the publication in order to satisfy the privacy preferences of all users affected. This whole process is the automatic generation of explanations to prevent privacy violations that we will present in this work.

3. Method

To properly generate explanations in our argumentation system, it is of the utmost importance to process the features extracted in a coherent way considering the argumentation framework. Computational arguments must be generated taking those features into account and finally, the explanations must be built in such a way that they can be interpreted by humans. It is important to emphasise that, since the implementation has been carried out in PESEDIA [1], the method proposed is not only compliant with the argumentation framework [15] requirements, but also with the social network functions and limitations.

3.1. Features

Four different features are considered by our proposed method to generate explanations: Privacy Preferences, Trust, Privacy Risk Score (PRS) and Sensitive Content. All those features are retrieved by the *feature extraction* module that works as an intermediary between the social network and the argumentation system. Two main sources can be identified when regarding the acquisition of the features: user preferences data and publication data.

3.1.1. Privacy Preferences

When sharing content in an online social network, the author must choose the target audience for each publication. For that purpose, privacy selectors such as the one depicted in Figure 1 are available in our educational social network. The options provided to the author as target audiences are the following: public, friends, collections and private. Public is the adequate option to share content with the whole network. With friends option it is possible to share the content only with the friends of the author. Collections are subsets of friends created by each user. Finally, the private option allow authors to keep a publication only visible to themselves.

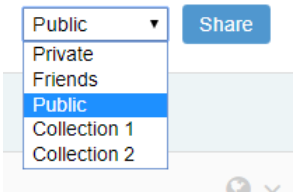

Figure 1. Privacy selector in our social network.

On the other hand, when creating a profile in the social network, each user must define the default target audience for their publications. Thus, the *feature extraction* module will retrieve both, users privacy preferences and the privacy configuration of the publication going to be shared.

3.1.2. Trust

When more users than the author are involved in a publication, it is important to respect every user data privacy to prevent multi party privacy conflicts. In PESEDIA, users can evaluate their friendship with a star based rating as depicted in Figure 2. This evaluation is retrieved by the *feature extraction* module, allowing us to quantify the existing trust between both users.

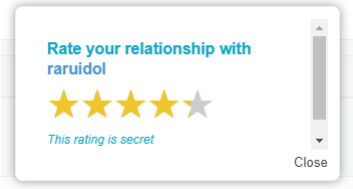

Figure 2. Relationship rating in our social network.

3.1.3. Privacy Risk Score

When sharing content in online social networks, it is hard to estimate the scope of a publication. The Privacy Risk Score [3] is a metric that allows us to obtain a value representing the risk of being read by unexpected users. Therefore, the *feature extraction* module will retrieve the computed PRS value for every publication going to be shared.

3.1.4. Sensitive Content

The content of the own publication is also an important feature to be considered when preventing privacy violations. Based in [7] work, we considered the following six categories of sensitive content: *location*, *medical*, *drug*, *personal*, *relatives* and *offensive*. Before sharing any publication, a text content analyser developed in the PESEDIA framework [2] is used to detect any of those categories of sensitive content. A six dimensional vector is generated by the analyser indicating the detected categories in the text. The *feature extraction* module retrieves the generated vector to feed the argumentation system with the data related to the sensitive content detected.

3.2. Argument Generation

Once all the features have been retrieved, the computational arguments are generated. Those arguments are defined by three parameters: the claim, the type and the support. The claim of an argument in this domain can have two perspectives, positive arguments in favour of sharing the publication or negative arguments against doing it. Secondly, each argument can belong to four different classes depending on the source of the features used to generate it: Privacy, Trust, Risk and Content arguments. Finally, the support is

a value computed from the extracted features that allow to both, quantify the individual strength of an argument and determine the claim of it.

Privacy arguments are generated from the privacy features. For this purpose, privacy options are represented with the following values: public (0), friends (0.5), collections (0.75) and private (1). Then, the support of the argument is computed as the difference between the author's default target audience and the publication privacy configuration. If the resulting value is negative, a privacy argument against making the publication will be generated, since the privacy configuration of the publication is less restrictive than the author's privacy preferences. Conversely, if the resulting value is not negative, a privacy argument in favour of making the publication will be generated.

Trust arguments are generated only when more than one user is tagged in a publication. The evaluation provided by users when adding friends (i.e., a 0-1 ranged value computed from the star based rating) is taken into account as the support for these arguments. We fixed a threshold of 4/5 stars (i.e., 0.8 in the numerical value scale) to start considering that trust may be enough to share some content without previously consulting. Therefore, a positive trust argument will be generated for each tagged user that has rated the author with 4/5 or 5/5 stars when evaluating the friendship. Conversely, if less stars are given from the tagged users to the author, trust arguments against making the publication will be generated.

Risk arguments are generated with the Privacy Risk Score as their own support. A threshold has been defined to discriminate between safe publications and publications that may cause a privacy violation. Due to the nature of this metric, we defined that threshold in the value 0.2, considering all publications with higher PRS dangerous for the author. Therefore, an argument against doing the publication will be generated if the threshold is surpassed. On the other side, an argument in favour of doing the publication will be generated if the threshold is not surpassed.

Content arguments are always generated against making the publication if any type of sensitive content is detected. The preference value of the user towards the specific type of content detected is used as the support of the generated argument. That value is modelled as follows,

$$v(t) = max(1 - \frac{n_t}{N}, \varepsilon) \tag{1}$$

where n_t is the number of publications containing some specific type of content t and N is the total amount of publications made by the user. This way, if any type of content predominates in some user publications, we can give priority to other types of content (e.g., a user whose account is medical content based will not be concerned about sharing medical content, but it may be about sharing other types of content). Initially, every user has the same value (1) towards each type of content. To smooth the initial decrements, we use ε to make sure the preference value is always higher than zero.

Once all the arguments are generated, an aggregation of the scores of every argument is done to compute the acceptable set as explained in [15]. Using complete semantics, only one of the extensions of arguments either in favour or against sharing the publication can be accepted. In the case of accepting the extension of arguments positioned against making the publication, automatically generated explanations will be shown to authors.

3.3. Explanations

In this method, we propose the use of templates to generate the explanations from the set of acceptable arguments. Table 1 contains how each type of computational argument is translated into natural language in order to be correctly interpreted by human users. Therefore, as many explanations as there are computational arguments in the acceptable set, will be generated and displayed to the author. In the next section, we provide an example of the automatic generation of explanations in the PESEDIA network.

Type of Argument	**Explanation Generated**
Privacy	The publication is going to be read by... (no one., your friends., a collection of friends., all the users.)
Trust	Some of the people you mention might get upset.
Risk	Your publication may be read by unknown people.
Content(Location)	You can be revealing information about where you are or where you're going.
Content(Medical)	You may be publishing private medical information.
Content(Drugs)	People might think you're on drugs/alcohol.
Content(Personal)	You could be publishing sensitive personal data.
Content(Relatives)	You could be making public information related to family or friends.
Content(Offensive)	Your publication might offend the people who read it.

Table 1. Explanations automatically generated by our system.

4. Case Study

A brief example of the proposed method for automatically generating explanations is depicted in both Figure 3 and Figure 5. For this example, we have considered a user profile publicly sharing a message that contains several privacy violations (Figure 3).

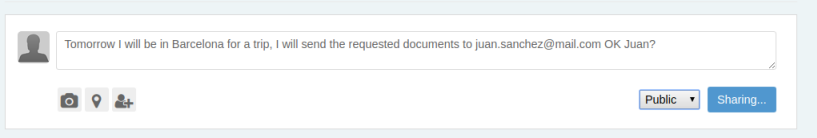

Figure 3. Example of publication triggering the argumentation system.

The author has his/her default target audience configured with the public option. Additionally, his tagged friend Juan, has rated his relation with the author with only three stars. From all the retrieved features (Table 2) in this scenario, a computational argument in favour of making the publication (i.e., Privacy) and four different computational arguments against sharing the content (i.e., Risk, Trust, Content(Location) and Content(Personal)) will be generated by the argumentation system. The corresponding

	User Preferences Data	**Publication Data**
Privacy	Public	Public
Trust	-	0.6 (Juan to Author)
Risk	-	0.75
Content(Location)	0.4	✓
Content(Medical)	1	-
Content(Drugs)	0.9	-
Content(Personal)	0.8	✓
Content(Relatives)	0.9	-
Content(Offensive)	1	-

Table 2. Features retrieved by the *feature extraction* module in the case study scenario.

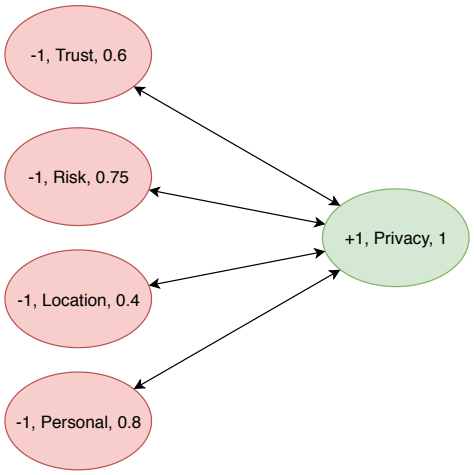

Figure 4. Argumentation graph generated by the *argument generation* module in the case study scenario.

argumentation graph generated by the argumentation system in this scenario is depicted in Figure 4.

With this example it is possible to observe the important difference between Privacy and Risk arguments. Although the author has configured the publication with a privacy policy coherent with his/her preferences, a Risk argument against doing the publication is generated since it is hard to predict to which audience the publication is going to reach. Once the inner procedure of the argumentation system has determined the set of acceptable arguments (against doing the publication), in order to prevent the potential privacy violation, explanations are automatically generated (Figure 5). The author can respond to the generated explanations either accepting them and modifying the original content, asking for more reasons, or else, ignoring them. Therefore, the final decision still relies on the user, but at least it will be made with a clear view on its potential consequences.

Figure 5. Example of automatic generated explanations. An explanation generated from a personal content argument (left) and an explanation generated from a trust argument (right) are depicted.

5. Discussion

In this work, we have proposed a method to automatically generate explanations to prevent privacy violations. For this purpose, an educational social network and an underlying argumentation framework for social networks have been used. The method proposed uses all the available information from the social network and the mechanisms provided by the argumentation framework to properly generate explanations for preventing a potential privacy violation. A case study has also been presented to illustrate the proposed method when facing a real situation.

We foresee several improvements as future work. In the current method, except for privacy arguments (where the audience is considered to generate the explanation), all the explanations are the same without taking into account anything else than the type of the argument. It would be interesting to explore the possibility of generating different explanations for the same type of arguments depending on the *strength* of them. Another interesting future work task would be to consider not only text, but also image data shared in the network, since data protection rules are even stricter on videos and photos, and many privacy violations are made by these means. Finally, to complement the improvements presented above, we are also considering adding a new type of argument based on the potential legal consequences of sharing some specific content. This new type of argument and its subsequent explanation will be useful in trying to persuade the author of a publication to modify it, warning of the possible legal consequences of sharing the publication.

Acknowledgements

This work is partially supported by the Spanish Government project TIN2017-89156-R, the FPI grant BES-2015-074498, and the Valencian Government project PROMETEO/2018/002.

References

[1] J Alemany. Pesedia. red social para concienciar en privacidad. 2016.

[2] J Alemany, E del Val, and Ana García-Fornes. Empowering users regarding the sensitivity of their data in social networks through nudge mechanisms. In *Proceedings of the 53rd Hawaii International Conference on System Sciences (in press)*, 2020.

[3] J Alemany, Elena del Val, J Alberola, and Ana García-Fornes. Estimation of privacy risk through centrality metrics. *Future Generation Computer Systems*, 82:63–76, 2018.

[4] Michał ARASZKIEWICZ and Grzegorz J NALEPA. Explainability of formal models of argumentation applied to legal domain.

[5] Trevor Bench-Capon. Before and after dung: Argumentation in ai and law. *Argument Computation*, pages 1–18, 11 2019.

[6] Trevor Bench-Capon, Henry Prakken, and Giovanni Sartor. Argumentation in legal reasoning. In *Argumentation in artificial intelligence*, pages 363–382. Springer, 2009.

[7] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pages 35–46. ACM, 2014.

[8] Cash Costello. *Elgg 1.8 social networking*. Packt Publishing Ltd, 2012.

[9] eMarketer. Number of social network users worldwide from 2010 to 2021 (in billions) Statista. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, 2019.

[10] Ricard L Fogues, Pradeep K Murukannaiah, Jose M Such, and Munindar P Singh. Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(1):5, 2017.

[11] Ricard L Fogues, Pradeep Murukanniah, Jose M Such, Agustin Espinosa, Ana Garcia-Fornes, and Munindar Singh. Argumentation for multi-party privacy management. 2015.

[12] Nadin Kökciyan, Nefise Yaglikci, and Pinar Yolum. An argumentation approach for resolving privacy disputes in online social networks. *ACM Transactions on Internet Technology (TOIT)*, 17(3):27, 2017.

[13] Yavuz Mester, Nadin Kökciyan, and Pınar Yolum. Negotiating privacy constraints in online social networks. In *International Workshop on Multiagent Foundations of Social Computing*, pages 112–129. Springer, 2015.

[14] Primal Pappachan, Roberto Yus, Prajit Kumar Das, Tim Finin, Eduardo Mena, Anupam Joshi, et al. A semantic context-aware privacy model for faceblock. In *Second International Workshop on Society, Privacy and the Semantic Web-Policy and Technology (PrivOn 2014), Riva del Garda (Italy)*, 2014.

[15] Ramon Ruiz-Dolz, Stella Heras, J Alemany, and Ana García-Fornes. Towards an argumentation system for assisting users with privacy management in online social networks. In *Proceedings of the 19th Workshop on Computational Models of Natural Argument.*, pages 17–28, 2019.

[16] Katharine Sarikakis and Lisa Winter. Social media users' legal consciousness about privacy. *Social Media+ Society*, 3(1):2056305117695325, 2017.

[17] Elizabeth I Sklar and Mohammad Q Azhar. Explanation through argumentation. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pages 277–285. ACM, 2018.

[18] Anna C Squicciarini, Federica Paci, and Smitha Sundareswaran. Prima: a comprehensive approach to privacy protection in social network sites. *annals of telecommunications-annales des télécommunications*, 69(1-2):21–36, 2014.

[19] Anna C Squicciarini, Heng Xu, and Xiaolong Zhang. Cope: Enabling collaborative privacy management in online social networks. *Journal of the American Society for Information Science and Technology*, 62(3):521–534, 2011.

[20] Kurt Thomas, Chris Grier, and David M Nicol. unfriendly: Multi-party privacy risks in social networks. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 236–252. Springer, 2010.

[21] 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 seventh symposium on usable privacy and security*, page 10. ACM, 2011.

[22] Ryan Wishart, Domenico Corapi, Srdjan Marinovic, and Morris Sloman. Collaborative privacy policy authoring in a social networking context. In *2010 IEEE International Symposium on Policies for Distributed Systems and Networks*, pages 1–8. IEEE, 2010.