Privacy Risk Assessment of Textual Publications in Social Networks

David Sánchez and Alexandre Viejo
Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili,
Av. Països Catalans, 26. 43007 Tarragona, Spain

Keywords: Privacy, Data Semantics, Information Theory, Social Networks.

Abstract: Recent studies have warned that, in Social Networks, users usually publish sensitive data that can be exploited by dishonest parties. Some mechanisms to preserve the privacy of the users of social networks have been proposed (i.e. controlling who can access to a certain published data); however, a still unsolved problem is the lack of proposals that enable the users to be aware of the sensitivity of the contents they publish. This situation is especially true in the case of unstructured textual publications (i.e., wall posts, tweets, etc.). These elements are considered to be particularly dangerous from the privacy point of view due to their dynamism and high informativeness. To tackle this problem, in this paper we present an automatic method to assess the sensitivity of the user's textual publications according to her privacy requirements towards the other users in the social network. In this manner, users can have a clear picture of the privacy risks inherent to their publications and can take the appropriate countermeasures to mitigate them. The feasibility of the method is studied in a highly sensitive social network: *PatientsLikeMe*.

1 INTRODUCTION

Social networks are virtual platforms where millions of users publish and read huge quantities of information every day. Even though a large part of this user-generated content may be considered innocuous, studies such as (Consumer Reports National Research Center, 2010) state that more than half of the users of social networks share private data. This dangerous behaviour jeopardises their privacy and, in several cases, it is the result of their privacy unawareness (Wang et al., 2014).

Several studies such as (D'Arcy, 2011; Zhang et al., 2010) have shown that the sensitive data that is published in social networks can be used by dishonest parties to perform threatening activities such as phishing, bullying or stalking among others. Due to the proliferation of these practices, in the last years, the privacy concerns of the users have grown significantly. Specifically, social network operators have recognised this situation and they have reacted by implementing privacy settings that enable users to decide who can access certain contents such as profile attributes or unstructured textual publications (i.e., wall posts, tweets, etc.).

Regarding unstructured textual publications, these elements are very dynamic and highly informative and they usually express up-to-date personal opinions or user behaviours and, therefore, they are considered to be particularly risky from the privacy perspective.

In any case, a major problem is that social networks do not provide mechanisms that enable the users to be aware of the sensitivity of the contents they publish (Wang et al., 2014). Therefore, many users are not aware of the privacy risks that their publications may cause and/or find difficulties in defining effective privacy settings.

Even though social networks operators have not paid enough attention to this issue, the scientific community has acknowledged its relevance. Some works try to measure the privacy risks that arise when users publish profile attributes in social environments. However, assessing the privacy risks related to publishing unstructured textual messages has received a lot less attention.

1.1 Previous Work

The authors in (Becker and Chen, 2009) developed a tool that focuses on Facebook and infers the hidden profile attributes (i.e., age, country, political view, relationship status, etc.) of a user from the publicly published attributes of her friends in this social network. The total number of attributes that can be

Sanchez D. and Viejo A..
Privacy Risk Assessment of Textual Publications in Social Networks.
DOI: 10.5220/0005281202360241
In *Proceedings of the International Conference on Agents and Artificial Intelligence* (ICAART-2015), pages 236-241
ISBN: 978-989-758-073-4

gathered by third parties is used to provide a privacy score to the users. Moreover, this tool proposes some user actions that can help to mitigate the detected privacy risks, which are based on removing certain friends or controlling them.

Following the same idea, the proposal presented in (Talukder et al., 2010) tries to infer private profile attributes from her friends' profiles. In this case, the tool is assumed access private information from the friend profiles that are made available to friends only. This tool presents to the user a measure of her privacy, a ranking of her friends based on individual contributions to privacy leakage and self-sanitisation actions to lessen this leakage.

Measuring the privacy exposure of a user by her profile attributes have been also proposed in (Srivastava and Geethakumari, 2013; Wang et al., 2014; Liu and Terzi, 2010). These three schemes consider the visibility of the attributes in the user profile (i.e., how known a certain item becomes in the network) in order to compute their respective privacy measures and also their sensitivity.

Finally, the authors in (Akcora et al., 2012) follow a different approach which is based on providing a risk measure to help users in judging a stranger in a social network. The goal is to inform users about how much it might potentially be risky to establish a connection with the stranger. The proposed tool provides information about the similarity between the user and the stranger as well as the information that the user may get from the stranger's profile.

So far, all the proposals are designed to measure the privacy risks in social networks related to certain profile attributes of users. However, the privacy risks inherent to the publication of textual messages, which represent most of the published contents and are highly informative, have been ignored in related works.

This issue is acknowledged in (Srivastava and Geethakumari, 2013), where the authors compute a measure of privacy risk of user's publications. Nevertheless, the work is quite preliminary and limited because: i) publications are evaluated as a whole, and thus the user cannot be aware of the specific message or term that causes a highest privacy threat; ii) publications are linked to a predefined set of attribute profiles, but the linking process is not detailed; iii) the privacy risks are computed by measuring the information distribution within a subset of social networks users, which would hardly reflect the actual information distribution at a social scale; and iv) privacy score calculations are ad-hoc.

1.2 Contributions

In this paper, we propose a fully automatic mechanism to inform users of social networks about the privacy risks inherent to their publications (i.e. unstructured textual messages). Compared with related works, our proposal is able to identify the specific terms within a message which can cause privacy risk with respect to concrete readers. In this manner, social network users can have a clearer picture of the privacy threats of their publications and, thus, take the appropriate protection measures.

The proposed method automatically assesses which content to be published could be risky according to its *degree of sensitivity* and the *privacy requirements* of the user, which are defined a priori according to the different types of contacts of the user in the social network. To do so, the method implements an information theoretically assessment of the disclosure risk of the published data as a way to measure the amount of knowledge (i.e. semantics) disclosed by a user to her potential readers.

2 THE PROPOSED METHOD

First, we detail how the sensitivity of the published data can be automatically assessed according to the amount of disclosed semantics. Then, we explain how the privacy requirements of the user can be gathered in an intuitive way. Finally, we detail how we evaluate the privacy risks inherent to textual publications according to the type of reader.

2.1 Sensitivity Assessment

The cornerstone of the proposed method is the assessment of the sensitivity of the information published by the user. Given that we are dealing with textual data and that this data is understood by humans (i.e. content producers, readers and also potential attackers) according to their semantics, we require a mechanism that measures the amount of semantics disclosed by the presence of each textual term in the user's publications. Our assumption, which is coherent with current research on document protection (Abril et al., 2011; Sánchez et al., 2013a), is that sensitive terms are those providing a large amount of semantics, because these are the ones that disclose more knowledge to attackers. However, since semantics are an inherently human and qualitative feature, their measurement is not trivial. To tackle this problem, we adopt an information theoretic quantification of data semantics. The basic

idea is that the semantics encompassed by a term appearing in a context can be quantified by the *amount of information* it provides, that is, its *Information Content* (IC).

The notion of IC has been extensively used to quantify term semantics (Resnik, 1995; Sánchez et al., 2011). The IC of a term t and, thus, the semantics encompassed by t, is computed as the inverse of its probability of occurrence in corpora.

$$IC(t) = -\log_2 p(t) \quad (1)$$

Thus, general terms such as *disease* provide less information and, thus, are less sensitive than specialised ones such as *lung cancer*, because the former are probably referred in a discourse.

Ideally, if the corpora used to compute the IC is large and heterogeneous enough to reflect the information distribution at a social scale, IC values will be a faithful representation of term's semantics as they are understood and used by humans. To compute realistic term probabilities, we rely on the largest and most up-to-date electronic repository available: the Web. In fact, the Web is so large and heterogeneous that it is said to be a faithful representation of the information distribution at a social scale (Cilibrasi and Vitányi, 2006), an argument that has been supported by recent works focusing on privacy-protection (Chow et al., 2008; Sánchez et al., 2013b; Sánchez et al., 2013a; Sánchez et al., 2014), which considered the Web as a realistic proxy for social knowledge.

In order to compute term probabilities from the Web in an efficient manner, several authors (Sánchez et al., 2010; Turney, 2001) have used the *hit count* returned by a Web Search Engine when querying the term t. Thus, in our approach, IC is computed as follows:

$$IC_{web}(t) = -\log_2 \frac{hits(t)}{N} \quad (2)$$

where N is the number of web resources indexed by the web search engine.

2.2 Privacy Requirements

In order to raise the users' awareness of the privacy risks that their publications may cause, we need to capture such her notion of privacy. We refer to this as the *privacy requirements* of the user. These state the degree of trust, and hence, the desired level of knowledge/information disclosure for the different types of readers in the social network.

Coherently with the privacy settings implemented in most social networks (Carminati et al., 2009), the *privacy requirements* in our approach are defined for the different types of contacts that the user may have in the social network. These types of relationships are specified by the users themselves when they add a new contact to their list of friends. Any unclassified entity will be classified in the lowest level of trust. Therefore, requirements are defined as a list of n of *privacy levels* that allow to classify the published data according to its degree of sensitivity and that are associated to each type of contact in the social network.

Then, by relying on the sensitivity assessment explained in the previous section, the published data is organised in a way that the least informative terms (i.e. those disclosing the least amount of semantics), which do not require any protection, are classified in the lowest privacy level L_0 (i.e. they can be accessed by external non-classified readers), whereas the most informative ones, which would only be accessed by fully trusted contacts, are classified in the highest privacy level L_{n-1}. The main idea is that the type of relationship (i.e. trust) that exists between the owner of the data and each type of reader defines the maximum knowledge that the latter can obtain from the former's publications. In this manner, we can identify privacy risks when the sensitivity of data is too high for a type of reader, so that the user may take countermeasures to minimise such risk.

In order to define the specific *privacy requirements*, the user is asked to define the maximum knowledge that would be disclosed to each privacy level/type of relationship. To make this process straightforward, and inspired by the usual privacy settings implemented in available social networks, these requirements are automatically setup by answering a set of predefined questions about sensitive topics for each type of contact. Specifically, the system presents a set of questions related to different sensitive topics (e.g. religion, race, sexuality, medical history, etc.) and, for each one, the user has to decide the maximum knowledge that, at most, the readers belonging to each type of relationship/privacy level can obtain.

Questions can be inspired in current legislations on data privacy (e.g. EU Data Protection Directive (The European Parliament and the Council of the EU, 1995), US federal laws on medical data (Health Privacy Project, 2013), HIPAA (Department of Health and Human Services, 2000) etc.), and should cover each of the topics that are categorised as private (e.g. religion, sexual orientation, race, census data, locations, sensitive diseases, etc.).

Once questions have been answered, the informativeness of the answers for each type of

contact/privacy level L_i (computed as in eq. (2)) is used as threshold T_{Li} for that level. This threshold will be used to determine the level to which a certain textual term t published by the user (with $IC(t)$) belongs and, thus, warn the user about the potential privacy risks. If different questions about several sensitive topics are performed to the user, the answer with the lowest informativeness for each privacy level will be used as threshold.

2.3 Evaluation of Privacy Risks

To evaluate the privacy risk of a user publication before making it accessible to the readers, the system takes the user's message and her *privacy requirements*. Then, it performs several linguistic analyses to extract potentially risky terms, whose sensitivity is evaluated according to their informativeness (eq. (2)) with regard to the thresholds of each privacy level. Then an assessment of the privacy risks of the terms in the publication is made for each type of contact, so that the user can take appropriate countermeasures, such as restricting the access to the publication or replacing too sensitive terms by less detailed data.

Due to the fact that the textual messages published in a social network usually lack a regular structure, we use several natural language processing tools to detect sensitive terms. Specifically, because sensitive terms are mostly concepts or instances and these are referred in text by means of noun phrases (NPs), the system focuses on the detection of NPs. NPs are detected by means of several *natural language tools* which perform sentence detection, tokenisation (i.e. word detection, including contraction separation), part-of-speech tagging (POS) and syntactic parsing of text.

Each NP, which we refer generically as term t, is then classified in a privacy level according to its level of sensitivity according to the privacy requirements. Finally, if necessary, the user is warned about the privacy risks that some of the terms to be published may cause towards some of her contacts. The process is next described formally.

Assuming *privacy requirements* with n levels $\{L_0,...,L_{n-1}\}$ and their corresponding n thresholds $\{T_{L0},...,T_{Ln-1}\}$, for each term t in a message m do:

- If $T_{L0} > IC(t)$, t is not sensitive for any type of contact because it provides less information than such allowed in the least restrictive threshold T_{L0}. It can be published as is.
- If $T_{L1} > IC(t) \geq T_{L0}$, t is sensitive for readers in L_0 because it provides more information than the threshold for L_0. The user should restrict the access to m for external contacts.
- If $T_{Li+1} > IC(t) \geq T_{Li}$, t is sensitive for readers in L_i or below. Thus, the user should restrict the access for users belonging to L_i or below.
- If $IC(t) \geq T_{Ln-1}$, t is sensitive for all the contacts of the user. Thus, the user should either not publish the message or replace the sensitive term(s) by less detailed data.

3 EMPIRICAL STUDY

In this section we discuss and illustrate the applicability of the proposed method in a social network characterised by the sensitivity of the user's publications: *PatientsLikeMe*. This social network is devoted to share information about users' conditions in order to give and receive feedback from other patients and from the medical community.

To show the applicability of our method, we will simulate a user U. First, the *privacy requirements* of U will be defined. Then, the behaviour of our method will be illustrated by randomly picking up a *real* textual publication from *PatientsLikeMe*. After that, the message will be analysed to detect privacy risks according to the type of contacts of U.

3.1 Defining the Privacy Requirements

For *PatientsLikeMe*, the following three types of users, sorted by their level of trust, are considered: "clinician/researcher", "follower" and "regular user". The first one is a healthcare professional that uses the data published in the social network for healing/research purposes. A "follower" of a user U is any user who decides to follow and is accepted by U. Finally, a "regular user" is any unclassified user who owns an account, which thus belongs to lowest privacy level L_0.

In order to define the privacy requirements for each type of contact, several questions are asked to the user. Due to space limitations, we will illustrate this process with one question referring to the *condition* of the user, which is shown in Table 1. To minimise errors, predefined answers with different degrees of informativeness are given. Note that there could be as many possible answers as desired, and that the same level of information disclosure could be associated to different contacts; in this latter case, the thresholds of different levels will be the same.

According to the answers of the user U shown in Table 1, the privacy requirements are set by computing the threshold T_{Li} for each privacy level L_i, which are computed as the IC of the answers. In the

sample questionnaire, thresholds can be directly defined according to the condition that the user U defines in her profile and the answers to the questions with regard to that condition. Let us assume that U claims to have "HIV" (which is sensitive according to (Health Privacy Project, 2013)). According to the answers, L_2 readers can know everything, thus T_{L2} is infinite; L_1 readers can know at most her sensitive condition, thus, T_{L1} is computed according to the informativeness of HIV: $T_{L1} = IC(\text{"HIV"}) = -\log_2(47E6/17E9) = 8,5$; finally, L_0 readers can only know that the user has a condition, but not the specific type, thus, $T_{L0} = IC(\text{"Condition"}) = -\log_2(126E6/17E9) = 7,1$.

In the above calculations, IC is computed using eq. (2) and Bing as the Web Search Engine to retrieve the number of hits. The number of web resources indexed by Bing is set to 17 billions, as estimated in http://worldwidewebsize.com.

Table 1: Sample questionnaire and simulated answers.

With regard to your *condition*, select the maximum knowledge that you are willing to disclose in your messages for each type of contact in *PatientsLikeMe:*
• Clinician/Researcher: (Level L_2) ☒ can know everything about your condition ☐ can know your condition but no specific details ☐ can just know that you suffer from a condition
• Follower: (Level L_1) ☐ can know everything about your condition ☒ can know your condition but no specific details ☐ can just know that you suffer from a condition
• Regular user: (Level L_0) ☐ can know everything about your condition ☐ can know your condition but no specific details ☒ can just know that you suffer from a condition

3.2 Assessing Privacy Risks

For illustrative purposes, let us assume that U wants to publish the message shown in the first row of Table 2, which is a random message gathered from the social network and published by a user with HIV. First, the message is syntactically analysed to detect all the terms (noun phrases) that may refer to sensitive concepts. Then, according to the *privacy requirements* of U and the IC of the detected terms, the latter are classified in each privacy level and, thus, the sensitive terms for each type of reader are identified, as shown in the last rows of Table 2.

By looking at Table 2 and recalling the privacy requirements of U, we have that L_2 readers (*clinicians/researchers*) would have access to all the details (i.e. $T_{L3} = \infty$); thus, since $IC(t_j) < T_{L1}$ for all terms t_j, none of such terms would cause a privacy risk. On the other hand, L_1 readers would be able to learn that the user has HIV and thus, $T_{L2} = IC(\text{"HIV"}) = -\log_2(47E6/17E9) = 8,5$; however, more specific terms like *Candidiasis* or *Tuberculosis*, whose $IC(t_j) > T_{L2}$ (i.e. $IC(\text{"Candidiasis"}) = 11,3$ and $IC(\text{"Tuberculosis"}) = 10,2$) should be restricted. The same process is also applied for readers in L_0. The system thus outputs a set of terms that cause privacy risks for each type of reader, so that the user may decide whether to restrict the access to the message and/or to remove or replace those terms by less detailed information.

Table 2: Sample message published in *PatientsLikeMe* and subsequent analyses of its content.

Analysis	*Result*
Original message	I guess was infected with HIV in 2012. In May 2013, I had 7 days of fever, I think it was a sinusitis. The fever came back in June 2013. Suddenly appeared Candidiasis at my throat. The doctor asked how many antibiotics I had taken. So he asked HIV testing and it results positive. I got sick very fast, because was diagnosed Tuberculosis. I was admitted to a hospital because of a thrombosis.
Sensitive terms for L_2 ($T_{L2}=\infty$)	None
Sensitive terms for L_1 ($T_{L1}=IC(HIV)$)	Sinusitis, Candidiasis, antibiotics, HIV testing, Tuberculosis, thrombosis
Sensitive terms for L_0 ($T_{L0}=IC(Condition)$)	HIV, May 2013, 7 days, fever, sinusitis, fever, June 2013, Candidiasis, throat, antibiotics, HIV testing, Tuberculosis, thrombosis

4 CONCLUSIONS

Social networks do not provide mechanisms to make users aware of the privacy risks of their publications and, thus, to intuitively apply the provided privacy settings (i.e. restricting the access of certain users to specific content). The method proposed in this paper tackles this problem by providing an automatic assessment of the privacy risks of user's publications with respect to her contacts and her privacy requirements towards such contacts.

The proposal can be implemented either as a module to be added in existing social networks or as an external application running on the user's side, which performs an a priori assessment of privacy

risks so that the user may apply countermeasures. It is worth mentioning that users do not require technical knowledge to set their privacy requirements, since they can be intuitively defined by answering questions. Moreover, requirements can be made coherent with legislations on data privacy.

Future research will focus on dealing with the ambiguity (e.g., polysemy, synonymy, ellipsis) that usually appears when syntactically analysing text and when computing term's IC from raw web-scale statistics. Moreover, we also plan to engineer questionnaires that are appropriate for the scope of some of the most widely used social networks in order to conduct additional experiments.

ACKNOWLEDGEMENTS

This work was partly supported by the European Commission under FP7 project Inter-Trust, H2020 project CLARUS, by the Spanish Ministry of Science and Innovation (through projects CO-PRIVACY TIN2011-27076-C03-01, ICWT TIN2012-32757 and BallotNext IPT-2012-0603-430000) and by the Government of Catalonia (under grant 2014 SGR 537). This work was also made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation.

REFERENCES

Abril, D., Navarro-Arribas, G. & Torra, V. On the declassification of confidential documents. 8th International Conference on Modeling Decision for Artificial Intelligence, 2011. 235–246.

Akcora, C. G., Carminati, B. & Ferrari, E. Privacy in Social Networks: How Risky is Your Social Graph? IEEE 28th International Conference on Data Engineering, 2012. 9-19.

Becker, J. & Chen, H. Measuring Privacy Risk in Online Social Networks. Web 2.0 Security and Privacy Conference, 2009.

Carminati, B., Ferrari, E. & Perego, A. 2009. Enforcing access control in Web-based social networks. *ACM Transaction on Information and System Security,* 13(1)**,** pp 38.

Cilibrasi, R. L. & Vitányi, P. M. B. 2006. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering,* 19(3)**,** pp 370-383.

Consumer Reports National Research Center 2010. Annual state of the net survey 2010. *Consumer Reports,* 75(6)**,** pp 1.

Chow, R., Golle, P. & Staddon, J. Detecting Privacy Leacks Using Corpus-based Association Rules. 14th Conference on Knowledge Discovery and Data Mining, 2008. 893-901.

D'Arcy, J. 2011. Combating cyber bullying and technology's downside. *The Washington Post.*

Department of Health and Human Services. 2000. The health insurance portability and accountability act.

Health Privacy Project. 2013. *State Privacy Protections* [Online]. Available: https://www.cdt.org/issue/state-privacy-protections.

Liu, K. & Terzi, E. 2010. A Framework for Computing the Privacy Scores of Users in Online Social Networks. *ACM Transactions on Knowledge Discovery from Data,* 5(1)**,** pp 30.

Resnik, P. Using Information Content to Evalutate Semantic Similarity in a Taxonomy. 14th International Joint Conference on Artificial Intelligence, 1995. 448-453.

Sánchez, D., Batet, M. & Isern, D. 2011. Ontology-based Information Content computation. *Knowledge-Based Systems,* 24(2)**,** pp 297-303.

Sánchez, D., Batet, M., Valls, A. & Gibert, K. 2010. Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems,* 35(3)**,** pp 383-413.

Sánchez, D., Batet, M. & Viejo, A. 2013a. Automatic general-purpose sanitization of textual documents. *IEEE Transactions on Information Forensics and Security,* 8(6)**,** pp 853-862.

Sánchez, D., Batet, M. & Viejo, A. 2013b. Minimizing the disclosure risk of semantic correlations in document sanitization. *Information Sciences,* 249(1)**,** pp 110-123.

Sánchez, D., Batet, M. & Viejo, A. 2014. Utility-preserving sanitization of semantically correlated terms in textual documents. *Information Sciences,* 279(1)**,** pp 77–93.

Srivastava, A. & Geethakumari, G. Measuring Privacy Leaks in Online Social Networks. International Conference on Advances in Computing, Communications and Informatics, 2013.

Talukder, N., Ouzzani, M., Elmagarmid, A. K., Elmeleegy, H. & Yakout, M. Privometer: Privacy protection in social networks. IEEE 26th International Conference on Data Engineering Workshops, 2010.

The European Parliament and the Council of the EU. 1995. *Data Protection Directive 95/46/EC* [Online].

Turney, P. D. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. 12th European Conference on Machine Learning, ECML 2001, 2001. 491-502.

Wang, Y., Nepali, R. K. & Nikolai, J. Social network privacy measurement and simulation. International Conference on Computing, Networking and Communications, 2014. 802-806.

Zhang, C., Sun, J., Zhu, X. & Fang, Y. 2010. Privacy and security for online social networks: Challenges and opportunities. *IEEE Network,* 24(4)**,** pp 13-18.