Challenges in Detecting Privacy Revealing Information in Unstructured Text

Welderufael B. Tesfay, Jetzabel Serna, and Sebastian Pape

Deutsche Telekom Chair of Mobile Business and Multilateral Security, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt am Main, Germany
{Welderufael.Tesfay, Jetzabel.Serna, Sebastian.Pape}@m-chair.de

Abstract. This paper discusses the challenges in detecting privacy revealing information using ontologies, natural language processing and machine learning techniques. It reviews current definitions, and sketches problem levels towards identifying the main open challenges. Furthermore, it elicits that the current notion of personally identifiable information lacks robustness to be used in varying contexts and user perceptions, and shows the need to additionally consider privacy sensitive information.

Keywords: Privacy, personally identifiable information, privacy sensitive information, privacy revealing information, ontology, machine learning

1 Introduction

These days many of our daily activities leave tremendous digital traces in the internet. While these data can be useful to solve societal challenges in areas such as health, transportation; it also presents threats to our personal and societal sovereignty with regard to privacy [11]. In this regard, the research community has been proposing different legal and technical mechanisms to safeguard privacy [5, 1]. At the centre of most of these proposals is the idea of anonymising the identity of users and reducing the disclosure of personal data. In order to reduce the disclosure of privacy revealing information (PRI), personalised privacy protection tools for internet users need to be developed. In this regard, Shah and Manisha [17] recommend machine learning techniques to detect PRI which often has a certain pattern. Similarly, Caliskan-Islam et al. [9] showed that ontologies can also be useful to detect privacy revealing textual data. However, regardless of the fact that personally identifiable information (PII) is often used as benchmark for privacy prevention measures, there is still no clearly established notion of what they consist of. This is due to complexities and context dependencies since the question of which information makes a user personally identifiable also depends on the group considered. This makes it particularly challenging to research in PRI detection, especially those using ontology, natural language processing and classification algorithms. Furthermore, there is the notion of privacy sensitive information (PSI), which highly depends on the user's own perception. For instance, in many social media sites the user is not anonymous,

and thus should be careful not to reveal PSI. This is even more important if the user is participating in the social media site over a long time and PSI could be obtained by a simple correlation of her postings. But like PII, the notion of PSI is not only fuzzy by itself, it also depends on whom the user is sharing the information with. For example, in anonymous support-groups each member reveals privacy sensitive information (the purpose of the support-group) but does not want the knowledge to diffuse outside of this group. In real-world groups the user is not anonymous to other members. However, in online anonymous support groups, e.g. an online forum with pseudonyms, the user may keep her identity private. Then it is particularly important, that the user does not reveal PII when posting in the forum, because that would threaten her anonymity. This should hold again, if the user is member of the group over a long time.

This shows, that – depending on the scenario/use-case – PII or PSI may be revealed (un)intentionally and the user's privacy is at risk as soon as someone is able to gather the matching part. We elaborate on this topic in Sect. 2.

In summary this paper identifies the challenges in detecting PII and PSI taking into consideration current approaches so far. The scope is limited to only textual content created by users, e.g. postings on social media sites such as Reddit. Thus the following data sources are out of scope of this work:

- Any (meta-)data collected on the user during the communication (e.g. with a social media site). This includes IP-address or fingerprinting the user's client
- Any other media than text. E.g. pictures or video files, even though they may contain meta-data such as exchangeable image file format (Exif) data.
- Information from other persons or users

The remainder of the paper is organised as follows. Sect. 2 discusses the notion and relation of personally identifiable and privacy sensitive information. Sect. 3 structures the underlying problems. In Sect. 4 related work is given and Sect. 5 focuses on open challenges. Sect. 6 concludes our work.

2 Terminology

A commonly used classification of privacy sensitiveness of data is based on whether the given data fulfils the definition of PII [3]. Identity theft and other privacy violations exploit these identifiers along with other background information of the target to compromise their privacy. However, privacy research hasn't yet developed an all-embracing definition for PII that will consider specific contexts. This induces a "grey area" between what kind of information should be considered PII and non-PII. Article 4 of the recently approved European General Data Protection Regulation (GDPR) [1] defines 'personal data' as any information relating to an *'identified'* or *identifiable natural person ('data subject')*. An identifiable natural person can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number or to one or more factors specific to the *physical, physiological, genetic, mental, economic, cultural or social identity* of that natural person [1]. Furthermore, Shilton [18]

stated that privacy decisions have multiple parts such as identity (who is interested in the data), granularity (how much the data reveals about the target) and time (for how long the data will be in use). Hasan et al. [8] also identified six different information types that could potentially identify a user: individual characteristics (age, name, gender etc), knowledge (information associated to specific knowledge in a given domain), interests (hobbies, sports, professional activities etc), goals (short term or long term users' wishes or intentions to achieve something in a given context), behavior (online activity, mobility patterns etc), and context information (spatio-temporal information).

In their recent work, Schwartz and Solove [16] developed a PII 2.0 model in which they propose a scale going from "no possibility of identification" to "individual can be clearly identified". This scale is further divided into three categories: (1) identified, (2) identifiable or (3) non-identifiable person or entity. In this context identified means that a specific person can be distinguished from others. In the second category the linkage to the specific person has not yet been made, but is likely to happen with additional information. The last category refers to data, which cannot be used to identify a person.

Thus far, we have seen how PII have been defined from technical and non-technical perspectives. However, as Narayanan and Shmatikov [13] have rightly mentioned, relying on PII for privacy risks analysis, such as re-identification of anonymised or pseudonymized information, at the age of big data is fallacious. For this reason we go back to the definition of personal data in the GDPR, which is defined as "any information relating to a [...] person". We argue that for analysing the user's privacy "any information" is too far-reaching and one should concentrate on sensitive data. However, analyzing privacy sensitiveness involves the users' own perception about the given attribute and goes beyond merely categorising data as PII or non-PII. Hence, we introduce the notion of privacy sensitive information (PSI), which is any information that, depending on the user's perception, has the consequence of revealing privacy of the individual. While PII and PSI have some commonalities, not all PII is necessarily PSI, and likewise not all PSI is PII. However, on the long run collections of PSI may result in PII. For further clarity, we introduce PRI, which is the superset of PII and PSI, and we recommend privacy research to consider PRI when dealing with privacy revelation detection and development of mitigating mechanisms.

Concluding, we note that – as already sketched in Sect. 1 – the critical path is linking a user's PSI with her PII. If only PSI or PII about a user is known, either the PSI cannot be linked to a user or there is no PSI about the identifiable user. This is in accordance with the definition in the GDPR.

3 Problem Statement

The task considered in this paper is to find PII and PSI in unstructured texts. We assume those texts are created by the user, e.g. as a posting in social media or a chat message. We do not consider structured texts, e.g. the fields in a form of a social media or chat user profile when the user is asked for certain attributes

such as age, gender, location. The task to check whether the input of a certain field belongs to the questioned category and is a truthful answer is fundamentally different than identifying relevant information in unstructured texts.

We have identified four different levels of difficulty depending on the amount of data considered.

1. *Identify PRI in an unstructured text*: Given an unstructured text, the challenge is to identify PII and PSI within the text. As already stated, the notion of PII depends on the context/domain and the notion of PSI additionally depends on the user perceptions, which information she classifies as sensitive.
2. *Identify PRI with historical information*: Given a number of unstructured texts, the challenge is to identify PII and PSI by analysing the set of (historical) texts. Since social media or chats usually involve a series of postings, in this level, historical information (previous unstructured texts) also needs to be considered. This results in the analysis of multiple texts, their timeline and the analysis of possible PII or PSI spread among them.
3. *Identify PRI with side information*: Given an unstructured text and publicly available information, the challenge is to identify PRI within the text by considering additional sources of information. PRI may be derived if the information contained in that text is combined with public knowledge, e.g. any kind of demographic data or public information such as Open Data or Wikipedia (cf. [14]). This is especially difficult due to the huge amount of available data and the large number of possible combinations of sources to infer PII or PSI from the text with (a subset of) the available side information.
4. *Identify PRI with side information and historical information*: Given a number of unstructured texts, the challenge is to identify PRI within the texts by considering additional publicly available information. This is the combination of available information in the problem levels 2 and 3.

The difficulty of the described levels forms a partial order. Obviously, levels 2 and 3 are more difficult than level 1 and level 4 is more difficult than levels 2 and 3. It is unclear if level 2 or 3 is more difficult than the other (cf. Fig. 1).

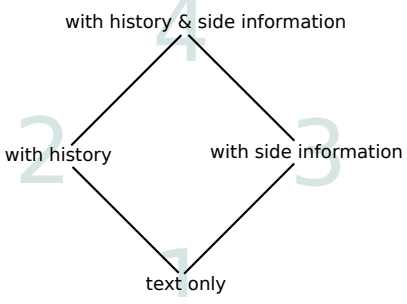

Fig. 1. Difficulty of different problem levels

4 Related Work

Detecting the presence and degree of sensitivity of private information is the first step towards empowering users with support in privacy decision making [9]. In this regard, Wang et al. [23] have introduced a tool to support social-science researchers, which performs a real-time analysis of unstructured short texts in order to extract conceptual associations. Authors focused on content generated in Twitter and implemented an association extraction module to determine the relevance of each word to the target keyword (e.g. privacy) using the real-time *pointwise mutual information* statistical association measure. However, the main outcome is not to identify privacy sensitive information, but instead to discover which concepts or topic users associate with privacy.

Mao et al. [12] presented an analysis of privacy leaks on Twitter. Their main contribution was to provide an initial understanding on what type of privacy information users reveal on their tweets. The authors limited the scope of their study to three privacy-related topics, namely, vacation plans, influence of alcohol and medical conditions. Furthermore, for each of those categories they built up a classifier in order to determine whether the tweet could be classified as privacy sensitive according to the detected content.

Jindal et al. [10] applied semi-supervised machine learning techniques to identify privacy sensitive data in medical texts. Their approach relied on information contained in the hierarchical structure of a large medical encyclopaedia. Following a similar direction, Caliskan-Islam et al. [9] proposed a privacy detective tool able to detect a broader range of privacy sensitive information. Authors combined a number of techniques, namely, topic modelling, named entity recognition, privacy ontology, and sentiment analysis in order to represent privacy features and trained a classifier based on Naive Bayes. They further analyzed Twitter users' time-line and computed a privacy scoring for each classifying them according to their sharing information behavior of privacy related information.

Gill et al. [7] developed a privacy dictionary able to distinguish between privacy and non-privacy content. The authors state that their privacy dictionary has six state categories. However, their categorisation lacks contextual and probabilistic rules to encompass more privacy related words. Zhang et al. [25] applied ontologies for rule-based privacy protection in pervasive computing environments. They used two properties (Data_is and and Disclose_when) to check for type of data and fulfilment of condition before data disclosure. However, the Data_is class consists of PII as defined in the Platform for Privacy Preferences Project (P3P) and was only extended by a location attribute. Therefore, this approach still lacks user's personal perception of privacy.

5 Challenges

In what follows, we illustrate the main challenges we have identified in the problem domain. Based on their mappings into the problem levels from Sect. 3, the challenges are further categorised into two, namely: general and specific ones.

General Challenges: these challenges regard all problem levels from Sect. 3.

Users' privacy perception - depending on different parameters such as educational background, previous privacy incidence/experience, perceived privacy risk, etc users have varying levels of concern and perceptions of privacy. Therefore, privacy detection systems built on the one-size-fits-all notion do not fairly address privacy choices of individuals. This is particularly challenging to rule-based ontologies and supervised machine learning approaches. Additionally, users' privacy perception may change over time.

Privacy paradox - users exhibit a complex and paradoxical dichotomy between their privacy (declarations) concerns and actual behaviour [24]. This is challenging to privacy disclosure analysis because on the one hand the decision involves whether to consider actual or declared behaviour, while on the other hand, prediction models can only grasp from the user activities, hence the actual behaviour, unable to accommodate the user's behavioural wishes.

Information privacy sensitiveness classification - when detecting PSI, for example illness, the degree of sensitiveness of a cold/flu might be substantially different from that of cancer [12]. Even though both attributes appear to fulfil privacy sensitiveness class, it is technically quite challenging to differentiate between the two. This is essentially related to the challenge of regarding the user's privacy perception except that this one focuses on the technicalities.

Context dependence - users privacy preferences are quite dependent on the context, e.g., a person can in a given situation be indifferent to privacy while the same person becomes very concerned about privacy in another situation [2]. If we consider the location data, a user may worry less about sharing it when in work place than when in a bar. In this case, ontologies and machine learning models can easily detect it as PII, but whether it is PSI or not strongly depends on the context.

Domain specific - there is a lack of comprehensive privacy dictionaries, ontologies, etc. which are not only context-aware [25], but have domain specific knowledge about PRI. Current ontologies are restricted for a specific target (e.g. legal compliance [4]) or a certain usecase (e.g. service orientation [6]).

Language issues - at the basic level text analysis requires proper sentence structures. Thus, text normalisation tools are needed before a proper analysis could be done. At the most advanced level, language variability also needs to be considered as languages are rich and there exist different ways to express certain meanings. Furthermore, most research works have focus on the English language and therefore, there is a lack of available resources

User preferences - users have different requirements on settings, e.g. there need to be different levels for experts or laymen. Getting this wrong, as well as mistakenly classifying PRI has a major effect on the usability, user acceptance and adoption of the provided tools.

Specific Challenges: these challenges map into the specific problem levels (cf. Sect. 3) as stated in each challenge below.

Data linkages resulting in PRI - by analysing multiple pieces of information new PRI could be found which are commonly not considered as such. With

the rise of publicly available information, especially open data (cf. [14]), linking seemingly harmless information with side information (cf. Sect. 3, problem level 3) could result in privacy breaches. This is one of the reasons why de-anonymization attacks (cf. Sweeney [21]) work.

Buildup of information - some information may not be privacy revealing if only a small quantity of data is considered, but may be sensitive in larger amounts. This holds for PII, e.g. revealing a location just one time might be safe, while you can build movement profiles from a continuous report of location data, e.g. by a mobile phone [20]. But this also holds for PSI, e.g. revealing only one base pair from your DNA might be safe while revealing a longer sequence might lead to implications [15]. The difficult question behind this challenge is, when is the aggregated amount of data privacy revealing? This is related to considering historical information (problem level 2).

PII and PSI may also be derived from other media, such as images or videos. Further work should consider the media itself, e.g. by using machine learning techniques [22, 19] to annotate/describe the content of the file. Additionally, further information within the files like exif-meta-data should be considered.

6 Summary and Conclusion

In this work, we investigated the challenges and opportunities in inferring privacy sensitive information from textual data using ontology, classification and natural language processing based techniques. While personal identifiable information and privacy sensitive information should be both considered, we showed fundamental differences between them. To structure the challenges in this area, we have defined different problem levels based on the information which needs to be regarded. A closer look at the challenges revealed that more research is needed, especially to consider specific privacy perceptions by different users when using the aforementioned techniques.

References

1. Eu general data protection regulation. Legal document, EU (2016), http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC, accessed July 2016
2. Acquisti, A., Brandimarte, L., Loewenstein, G.: Privacy and human behavior in the age of information. Science 347(6221), 509–514 (2015)
3. Callahan, M.: Us dhs handbook for safeguarding sensitive personally identifiable information. Washington, DC (2012)
4. Casellas, N., Nieto, J., Meroño, A., Roig, A., Torralba, S., Reyes, M., Casanovas, P.: Ontological semantics for data privacy compliance: The NEURONA project. In: Intelligent Information Privacy Management (2010)
5. Dwork, C.: Differential privacy. In: Automata, languages and programming, pp. 1–12. Springer (2006)

6. Garcia, D., Toledo, M.B.F., Capretz, M.A.M., Allison, D.S., Blair, G.S., Grace, P., Flores, C.: Towards a base ontology for privacy protection in service-oriented architecture. In: 2009 IEEE International Conference on Service-Oriented Computing and Applications (SOCA) (2009)
7. Gill, A.J., Vasalou, A., Papoutsi, C., Joinson, A.N.: Privacy dictionary: a linguistic taxonomy of privacy for content analysis. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 3227–3236. ACM (2011)
8. Hasan, O., Habegger, B., Brunie, L., Bennani, N., Damiani, E.: A discussion of privacy challenges in user profiling with big data techniques: The eexcess use case. In: Proceedings of the IEEE International Congress on Big Data (2013)
9. Islam, A.C., Walsh, J., Greenstadt, R.: 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 (2014)
10. Jindal, P., Gunter, C.A., Roth, D.: Detecting privacy-sensitive events in medical text. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '14 (2014)
11. Kum, H.C., Ahalt, S.: Privacy by design: understanding data access models for secondary data. American Medical Informatics Association (AMIA) Joint Summits on Translation Science and Clinical Research Informatics (2013)
12. Mao, H., Shuai, X., Kapadia, A.: Loose tweets: An analysis of privacy leaks on twitter. In: Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society (2011)
13. Narayanan, A., Shmatikov, V.: Myths and fallacies of personally identifiable information. Communications of the ACM 53(6), 24–26 (2010)
14. Pape, S., Serna-Olvera, J., Tesfay, W.: Why open data may threaten your privacy. In: Workshop on Privacy and Inference, co-located with KI (September 2015)
15. Roche, P.A., Annas, G.J.: Dna testing, banking, and genetic privacy. New England Journal of Medicine 355(6), 545–546 (2006)
16. Schwartz, P.M., Solove, D.J.: Pii 2.0: Privacy and a new approach to personal information. Privacy and Security Law Report (2012)
17. Shah, R., Valera, M.: Survey of sensitive information detection techniques: The need and usefulness of machine learning techniques
18. Shilton, K.: Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection. Commun. ACM (2009)
19. Siddiqui, A., Mishra, N., Verma, J.S.: Article: A survey on automatic image annotation and retrieval. International Journal of Computer Applications (2015)
20. Spitz, M.: Tell-all telephone. `http://www.zeit.de/datenschutz/malte-spitz-data-retention` (2010)
21. Sweeney, L.: Simple demographics often identify people uniquely. Tech. rep., Carnegie Mellon University (2000), data Privacy Working Paper 3
22. Wang, F.: A survey on automatic image annotation and trends of the new age. Procedia Engineering (2011)
23. Wang, Z., Quercia, D., Séaghdha, D.O.: Reading tweeting minds: Real-time analysis of short text for computational social science. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media (2013)
24. Young, A.L., Quan-Haase, A.: Privacy protection strategies on facebook: The internet privacy paradox revisited. Information, Communication & Society 16(4), 479–500 (2013)
25. Zhang, N.J., Todd, C.: A privacy agent in context-aware ubiquitous computing environments. In: IFIP International Conference on Communications and Multimedia Security. Springer (2006)