Increased Information Leakage from Text

Sicong Zhang, Hui Yang, Lisa Singh

Department of Computer Science
Georgetown University
37th and O Street, NW, Washington, DC, 20057
sz303@georgetown.edu, {huiyang,singh}@cs.georgetown.edu

ABSTRACT

The enormous data sharing and data availability on the Internet provides opportunities for new services tailored to extract, search, aggregate, and mine data in meaningful ways, At the same time, it poses challenges with regards to data privacy. This paper offers insight into this problem, focusing on current relevant research and potential areas of synergy between the information retrieval community and the privacy community. We then analyze example publicly shared data and discuss the types of data that can be extracted, the methods used for extracting them, and the implications for individuals who share personal information.

1. MOTIVATION

We are living in an era of enormous data sharing and data availability on the Internet. This data pervasiveness has resulted in the emergence of services tailored to extract, search, aggregate, and mine data in meaningful ways. On the one hand, this is a boon for information retrieval researchers. The flow of large volumes of textual data offers the perfect playground for developing new search and retrieval algorithms. On the other hand, the sharing of large amounts of data, some of which are sensitive, presents challenges with regards to data privacy. First, because of privacy laws, these data are not always available to researchers. Second, even if the data are readily available, what are the ethics of using data without explicit consent from the data owner? We understand the need for consent for private data, but what about public data? While legal, is it ethical? It is unclear about how the public feels about the sharing of their search results and it is even more unclear what researchers should do to mitigate potential concerns. What techniques exist to efficiently and effectively anonymize the data so that researchers can still work on traditional information retrieval tasks using personalized information? This dilemma leads us to the main question discussed in this paper. How do we protect the privacy of individuals while accessing and gathering this data to improve information retrieval algorithms and methods?

2. PRIVACY RESEARCH AND DIRECTIONS

There are a number of privacy models that have been proposed in the literature and are relevant to our community. Here we present some of the most relevant.

Leakage Across Social Network Sites: Understanding the data that users are willing to share and the level of sensitivity associated with it is a growing area of research. More specifically, a number of researchers are investigating how easy it is to link an individual across online social networks (OSNs) [3, 7]. In this problem, a user has accounts on multiple social networks. An adversary generally begins with a particular user's account information on one social network, i.e. the user's account id. The adversary then uses public data available on different OSNs to map profiles on these OSNs, exploiting the user's privacy with the additional knowledge gained. The research objective is to determine and quantify the level of leakage that exists for a large number of individuals across the OSNs. Existing research uses a number of different attributes to map individuals from one site to another, including account names, geo-location, post timestamp, social network connections, and structured demographic attributes to names a few. This form of record-linkage can be viewed as a search problem. Further, to date, studies are not incorporating knowledge from text in these analyses. Determining how to integrate textual knowledge into this process is another opportunity for researchers in the IR community.

Re-identification: Many companies and government agencies are either required by law or wish to release anonymized versions of their data. Unfortunately, given the amount of public data available, sometimes it is possible to un-anonymize the anonymized data. The goal of re-identification is to match or link anonymized personal data to publicly available data in order to determine sensitive data values of users in the anonymized data. In this threat model, the adversary has access to an anonymized data set and one or more public data sources containing a large sample of user data. The anonymized data usually contains sensitive data fields that if matched to an individual, would result in a significant privacy breach. The public data source generally contains *benign* data, i.e. fields that are not considered sensitive. Researchers have successfully shown that re-identification is possible with voting records, with data from OSNs, and with released medical data [1, 4, 6].

Data Publishing: When sensitive data needs to be released or published, companies must consider different approaches for maintaining confidentiality. Standard approaches include anonymization, adding noise to the data, binning data, and suppressing parts of the data. While many approaches have been proposed for relational data [8] and graph data [10], very few studies have investigated ways to anonymize textual data. The lack of research in this area is not surprising since text data has been readily available for years. Only in the last five to ten years have companies begun limiting access to their weblogs, search data, etc. These limitations, while understandable, limit the progress of research by limiting access to large, corporate data sets. Therefore, developing anonymization strategies that are optimized for text corpora is an important area of research.

Differential Privacy: Government agencies like the Census Bureau maintain statistical databases that by law need to be accessible to the public. In the case of the Census Bureau, their databases contain survey data results. These results need to be shared with the public without violating the privacy of the individuals who took the survey. While the Census Bureau uses many different techniques to maintain the privacy of individuals, many companies are developing techniques that follow the principles of differential privacy [2], a protocol that when used can provide the user with a clear probability of leakage when a single user is added or removed from the data set. Specifically, the protocol states that if an individual in the data set changes his/her data value a_i to any other allowable value a_j, then the difference between the privacy functions is smaller than a parameter ϵ. The strength of this approach is the provable privacy guarantees that many ad hoc methods lack. Examples and algorithms related to using differential privacy for sparse textual data sets is lacking and would be an important direction of research.

3. INFORMATION LEAKAGE IN TEXT

As a proof of concept, for the types of information that can be extracted using natural language processing (NLP) methods, we analyze a single LinkedIn profile. Before analyzing text fields, we mention that we can obtain up to 18 different types of data from each user including first name, last name, occupation, picture, education, location, skill, and company. These fields can be fairly unique if a single user shares all of them. Even more important is that once text analysis is conducted on the summary that many LinkedIn users provide, the amount of additional information that can be learned about the user can really increase. As an example, let us consider the following LinkedIn summary. Note that identifying attributes of this text passage have been replaced with dummy values to maintain privacy.

> *... at A Company, i support our clients by developing marketing and media plans, implementing social media campaigns, overseeing inbound marketing initiatives, and crafting the perfect pitches and press releases to secure news coverage. before joining the agency, i was involved in public relations, promotions and event planning for health care, government and education organizations. i assisted in planning and promoting fund raising events for a regional hospital foundation, and developed communication materials for various hospital programs. while in graduate school, i conducted research for B university on topics including nonprofits and crisis communication, public relations theory, and social media usage among nonprofits. i earned a ba in communication studies with a concentration in journalism from the C university, and a ms in public relations from D university. i wrote my masters thesis on corporate health diplomacy, where i reviewed the corporate social responsibility and business development efforts of pharmaceutical companies examined by international public relations literature ...*

For this user produced introduction paragraph in LinkedIn, some sensitive information can be extracted by the following procedure. First, we can apply part-of-speech tagging, shallow parsing, and named entity tagging on the paragraph. We obtain a list of noun phrases, such as “marketing and media plans", “social media campaigns", and “public relations". We also obtain a list of named entities, such as “A company", “C university", and “D university". Next, we compare the members in the noun phrase list as well as the members in the named entity list with a general ontology maintained by us. The ontology keeps a dictionary of terms that indicates various important aspects for a person. For instance, terms that indicate a person’s professional degree could be “ba", “ma", “phd", “mba", etc. Terms that indicate a person’s profession could be “sde", “accountant", ‘cpa", etc. The ontology is created by extracting terms and relations from Wikipedia hierarchies via an automatic approach [5]. Each major aspect of a person will have its own list of key terms. These aspects include profession, location, home address, gender, birthday, hobby, etc.

Then, if a free text paragraph contains many terms from the keyword term list for a certain aspect, we will determine whether the paragraph should be classified into one or more aspects. The category of the paragraph can thus be detected by text classification techniques [9]. The name of the resulting category can then be used to trigger the corresponding set of lexico-syntactic patterns. The purpose is to extract detailed information for a category. For instance, using the pattern “ms in NP from NE_university", we can obtain the Master of Science degree of a person in a university, if we keep NP in this example as a place holder. We can also obtain the name of the university where a person went to study Master of Science in “public relations". These relations extracted by the aspect-specific patterns can then be put into a relational database. Available techniques on data re-identification can be applied to find out the potential information leakage in the text.

4. FUTURE INFORMATION RETRIEVAL RESEARCH DIRECTIONS

Given the volume of blogs, tweets, and other textual personal data shared by users, it is time for our community to consider how data privacy will affect our field. We need to develop protocols for useful anonymized data sets that are non-invasive in terms of individual privacy. We need to understand what types of data can be learned by adversaries using textual data and better understand the sensitivity of learning these data. This is a time for the IR community to get involved in privacy research.

5. ACKNOWLEDGMENT

This research was supported by NSF grant CNS-1223825. Any opinions, findings, conclusions, or recommendations expressed in this paper are of the authors, and do not necessarily reflect those of the sponsor.

References

[1] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore Art Thou r3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography. In *ACM World Wide Web Conference (WWW)*, 2007.

[2] C. Dwork. Differential privacy. *Lecture Notes in Computer Science*, 4052:1–12, 2006.

[3] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting innocuous activity for correlating users across sites. In *ACM International Conference on World Wide Web (WWW)*, 2013.

[4] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *ACM Workshop on Privacy in the Electronic Society (WPES)*, 2005.

[5] D. Guan and H. Yang. Increasing stability of result organization for session search. In *ECIR*, pages 471–482, 2013.

[6] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *IEEE Symposium on Security and Privacy*, 2009.

[7] A. Ramachandran, L. Singh, E. Porter, and F. Nagle. Exploring Re-identification Risks in Public Domains. In *IEEE International Conference on Privacy, Security and Trust (PST)*, 2012.

[8] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, 1998.

[9] Y. Yang. An evaluation of statistical approaches to text categorization. *Inf. Retr.*, 1(1-2):69–90, May 1999. ISSN 1386-4564. doi: 10.1023/A:1009982220290. URL `http://dx.doi.org/10.1023/A:1009982220290`.

[10] B. Zhou, J. Pei, , and W. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations Newsletter*, 10(2), December 2008.