2016

An Attribtue-based Statistic Model for Privacy Impact Assessment

Yong Wang
Dakota State University

Jun Liu
Dakota State University

An Attribtue-based Statistic Model for Privacy Impact Assessment

Yong Wang
College of Computing
Dakota State University
Madison, SD 57042
yong.wang@dsu.edu

Jun Liu
College of Business and Information Systems
Dakota State University
Madison, SD 57042
jun.liu@dsu.edu

EXTENDED ABSTRACT

Personally Identifiable Information (PII) includes any information that can be used to distinguish or trace an individual's identity such as name, social security number, date and place of birth, mother's maiden name, or biometric records. It also includes other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information. PII is often the target of attacks, and loss of PII could result in identity theft. According to the U.S. Department of Justice, the average number of U.S. identity fraud victims annually is 11,571,900 [1]. The total financial loss attributed to identity theft in 2013 was $21 billion dollars, compared to $13.2 billion total loss in 2010 [1].

A. Introduction

PII is essential in protecting data security and privacy. HIPAA defines 18 identifiers that might be used to identify an individual as Protected Health Information (PHI) and requires the information must be protected in any form or medium. However, there exist additional identifiers that can be used to link to an individual. What is PII? What does PII include? How does publicly available information (e.g., public records, social networks, search engines, etc.) affect privacy? These questions are important. However, the answers to these questions are vague.

First, more identifiers may exist and can be used to link to an individual. For example, Montjoye et al. found that human mobility traces are highly unique [2]. Using a test dataset where the location of an individual is specified hourly, 95% of the individuals can be uniquely identified using four spatiotemporal points. Second, since the correlations of personally identifiable information are not clear, it is also uncertain what data should be protected due to deep analytic techniques. Inference attacks exist and can be used to collect more indirect information from existing known data which may be linked to personal identity [3]. Moreover, non-sensitive data could be aggregated to reveal more sensitive information and cause identity theft [4], [5]. Third, advance techniques such as de-anonymization could be used to link anonymous data to personal identity, and many efforts have been conducted on privacy preserving to obscure PII in datasets [6], [7]. However, de-anonymization attacks have been found to be effective in re-identifying anonymous data [8], [9]. Fourth, numerous public accessible information is available via public records, social media and the Internet. Information may not be available before and is now accessible on the Internet. It is not clear how the publicly available information affects user privacy. Fifth, privacy is also evolving and the definition of privacy may change as technology advances. A dynamic approach to characterize privacy is desirable.

Our studies show that developing an accurate model for PII is a fundamental issue to resolve many challenges in privacy, such as privacy measurement, data loss assessment, policy making, etc. This paper proposes to develop an attribute-based statistic model for privacy exposure measurement and privacy impact assessment based on text mining and machine learning.

B. An Attribtue-based Statsitc Model for PII

The PII model includes three key components: privacy attributes, privacy sensitivities, and attribute correlations.

Privacy Attributes: We use an actor to refer an entity (e.g., people, organization, etc.) on the Internet. An actor has certain characteristics, such as name, address, phone number, etc., which are known as "attributes". Privacy attributes are the attributes which may affect privacy. Privacy attributes describe what privacy is and what it includes.

Privacy Sensitivities: Each attribute has an impact on privacy. This impact is referred to as a privacy impact factor. A privacy impact factor is a numerical value. We consider the privacy impact factor for full privacy disclosure as 1. An attribute's privacy impact factor is a ratio of its privacy impact to the full privacy disclosure. Thus, an attribute's privacy impact factor has a value between 0 and 1. Privacy sensitivities describe how an attribute affects privacy.

Attribute Correlations: Attribute correlations describe how attributes are related. There are two correlations which need to be further explored, i.e., inference information and aggregated information. Inference information is hidden information which can be derived from a known attribute. For example, location of high school indicates an individual's hometown and hometown may be related to personal preferences, such as sports, etc. Attributes also show group properties. For example, as found in [4], [5], 87% of Americans can be uniquely identified by five digit zip code, gender, and date of birth. However, none of these characteristics alone can significantly affect privacy.

Let A be an actor and we assume the PII model includes m privacy attributes. We use a_i to represent the i-th privacy attribute. Thus, $A(a_1, a_2, \dots, a_m)$ describes all attributes which may affect A's privacy. We use s_i to represent the privacy sensitivity of i-th attribute. Thus, we have $S(s_1, s_2, \dots, s_m)$ representing A 's privacy attribute sensitivities. Attribute

978-1-5090-2300-4/16 $31.00 © 2016 IEEE
DOI 10.1109/CTS.2016.117

correlations are manifested in a number of rules. Let r_i be a correlation rule and we use $R = \{r_1, r_2, \dots, r_n\}$ to represent attribute correlations. Thus, an attribute-based PII model is defined as including:

$$\begin{cases} \text{Attributes: } A(a_1, a_2, \dots, a_m) \\ \text{Sensitivities: } S(s_1, s_2, \dots, s_m) \\ \text{Coorelations: } R = \{r_1, r_2, \dots, r_n\} \end{cases}$$

The model defines what privacy is, how an attribute affects privacy, and how the attributes are related. It can help resolve many challenges in security and privacy. For example, using attributes and sensitivities, data loss can be assessed and risk can be analyzed. Using attribute correlations, potential sensitive data can be identified and removed in the de-identification process.

1) Attribtue Extraction

Attribute extraction is based on text mining and machine learning. Three data sources are used for text mining:

- Privacy documents such as privacy laws, regulations, directives, policies, instruction letters, etc.
- Online social network and website user profiles
- Web search engines such as google, Bing, etc.

A knowledge base is established to include the initial PII model. The initial privacy attribute set includes the 18 identifiers defined in the HIPPA document. A document filter will be used to identify these three types of data sources. Text mining will then be conducted on these documents. Term frequency (*tf*) and inverse document frequency (*idf*) will be calculated for each term in a document. A weight (*tf x idf*) is assigned to each term. A raw set of attributes could be extracted based on the weight ranking. Other term weighting methods also exist [10].

The raw set of attributes needs to be further analyzed. We use term association and term similarities to further justify if an attribute is a privacy attribute or a duplicate attribute. Association of each raw attribute and existing privacy attributes will be checked. The association is represented using a true/false matrix where $t_1, t_2, \dots, t_m$ are the known privacy attributes, $t_{m+1}, t_2, \dots, t_{m+p}$ are raw attributes, $D_1, D_2, \dots, D_n$ are mining documents, and $k_{ij} = 0 \; or \; 1 \; (1 \le i \le n, 1 \le j \le m + p)$. $k_{ij} = 1$ indicates $t_j \in D_i$ otherwise $k_{ij} = 0$. We define co-occurrence value between privacy attribute t_j $(1 \le j \le m)$ and term t_{m+l} $(1 \le l \le p)$ as

$$coo(t_j, t_{m+l}) = \sum_{i=1}^{n} k_{ij} * k_{i,m+l}$$

A new attribute might be a privacy attribute if it has a high occurrence with an existing privacy attribute.

Similarities will also be checked between the raw attributes and the known privacy attributes where d_{ij} is the weight of term t_j in the document D_i. A similarity value will be calculated between the new attribute t_{m+l} $(1 \le l \le p)$ and the existing known attribute t_j $(1 \le j \le m)$.

$$sim(t_j, t_{m+l}) = \sum_{i=1}^{n} d_{ij} * d_{i,m+l}$$

A raw attribute can be identified as a new privacy attribute if it is not similar with any existing privacy attributes.

Thresholds for co-occurrence and similarities can be derived and used for machine learning to automate the process. Once a raw attribute is confirmed to be a privacy attribute, the attribute is added to the knowledge base for future analysis. This approach is based on text mining and machine learning. It is dynamic and will be very useful to track privacy trends on the Internet.

2) Attribute Senstivity Evalaution

Sensitivity is a numerical value between 0 and 1 which indicates an attribute's impact on privacy. As discussed, a term weight (*tf x idf*) can be derived based on text mining. A normalized term weight between 0 and 1 will be used as an initial sensitivity value. Using a term weight as attribute sensitivity may have limitations. For example,

- The connection between term weight and privacy sensitivity needs to be justified.
- Using term weight as privacy sensitivities may violate attribute correlations.

These two limitations can be further improved using the second approach, sensitivity justification and adjustment based on discovered constraint rules.

Attribute sensitivities are not random value and they must obey constraint rules. We have observed that the following rules must be satisfied when assigning sensitivities to privacy attributes. Let a_1 and a_2 be two attributes and s_1 and s_2 be their sensitivities,

- If a_1 is privacy attribute and a_2 is not, $s_1 > s_2$;
- If a_1 is essential to a user than a_2, $s_1 > s_2$;
- If a_1 is used more frequently than a_2 in security incidents, $s_1 > s_2$;
- If a_1 can be inferred from a_2, $s_1 \le s_2$;

More rules might be discovered and used when we evaluate attribute sensitivities. These constraint rules will also be part of the knowledge base to guide the text mining and machine learning. We will use these rules to justify and adjust the sensitivity value assigned to an attribute [11].

3) Attribtue Correlation Revelation

Attributes are not independent. They may depend on each other. We are interested in capturing two particular kinds of correlations, inference information and aggregated information. Attribute correlations can be described by $R = \{r_1, r_2, \dots, r_n\}$ where r_i is a correlation rule such as

$$V_{i1} \overset{p_i}{\rightarrow} V_{i2},$$

where V_{i1} and V_{i1} are two subsets of the privacy attribute set and p_i $(0 \le p_i \le 1)$ is a probability to indicate how much information of V_{i2} can be learned from V_{i1}. For example, correlation rule $\{a_i\} \overset{p=1}{\longrightarrow} \{a_j\}$ describes an inference rule which indicates that attribute a_j can be inferred from attribute a_i. Correlation rule $\{a_i, a_j, a_k\} \overset{p=1}{\longrightarrow} \{a_t\}$ describes an aggregation rule which indicates that attribute a_t can be further decided by attributes a_i, a_j, and a_k together. The correlation rule can be

further represented as a function $h(V_{i1},V_{i2}) = p_i$ to simplify calculation.

Text mining approaches such as Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA) have been widely used in information retrieval to identify term relations [12], [13]. Text mining can also be used to look for and identify privacy attribute relations, such as association and co-occurrence. Associated attributes and attributes in a cluster may indicate attribute correlations (inference information or aggregated information). Thus, a raw attribute correlation can be derived. A raw correlation may not be true or reasonable. It needs to be further verified and justified. Publicly available data, such as social network user data and data available on the Internet, will be used to justify attributes, sensitivities, and correlations. A justified attribute correlation will then be added to the PII model knowledge base.

Attribute correlations can be further expanded using approaches in machine learning. An initial correlation knowledge base will be established based on the PII model. The correlation knowledge base will be further expanded using the new correlations discovered in text mining and machine learning. Reasoning is then used to expand the correlation knowledge base using existing inference information and aggregated information. For example, deductive reasoning can be conducted on the existing correlations, $\{a_1\} \xrightarrow{p_{12}} \{a_2\}$ and $\{a_2\} \xrightarrow{p_{23}} \{a_3\}$. New correlation $\{a_1\} \xrightarrow{p_{13}} \{a_3\}$ can be deduced and added to the knowledge base.

C. Privacy Impact Assessment

The PII model includes privacy attributes, attribute sensitivities, and attribute correlations. To measure privacy exposure, privacy measurement functions and attribute visibilities are required. Our previous works on privacy measurement proposed three functions for privacy measurement: weighted privacy measurement function, maximum privacy measurement function, and composite private measurement function [14]. Correspondingly, three privacy indexes are defined: weighted-privacy index, maximum-privacy index, and composite-privacy index [14]. Using the PII model and privacy indexes, privacy exposure can be measured and privacy impact can be assessed on the Internet.

D. Summary

An attribute based statistic model is proposed to measure privacy exposure and assess privacy impact based on personal identifiable information. The model includes three key components, i.e., privacy attributes, privacy sensitivities, and attribute correlations. The approaches used to develop the model are based on text mining and machine learning. It is different than the existing approaches used in natural language processing and the approaches used in studying contextual privacy. Natural language processing, e.g., SemEval [15], can determine the sense of the word 'privacy' in a context. However, it does not address attribute sensitivities and correlations. Contextual privacy targets to protect privacy in a context [16] based on the assumption that we know privacy and have clear definition of privacy. However, such definition and model for privacy is not available in real practice.

REFERENCES

[1] Javelin Strategy & Research, "2013 Identity Fraud Report: Data Breaches Becoming a Treasure Trove for Fraudsters," 2013.

[2] Y.-A. de Montjoye, C. a Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility.," *Sci. Rep.*, vol. 3, p. 1376, 2013.

[3] J. Tang, T. Lou, and J. Kleinberg, "Inferring Social Ties across Heterogeneous Networks," *WSDM '12 Proc. fifth ACM Int. Conf. Web search data Min.*, pp. 743–752, 2012.

[4] P. Golle, "Revisiting the uniqueness of simple demographics in the US population," in *Proceedings of the 5th ACM workshop on Privacy in electronic society*, 2006, pp. 77–80.

[5] L. Sweeney, *Uniqueness of Simple Demographics in the U.S. Population, LIDAP-WP4*. Carnegie Mellon University, Laboratory for International Data Privacy, 2000.

[6] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-Diversity: Privacy beyond k-anonymity," in *Proceedings - International Conference on Data Engineering*, 2006, vol. 2006, p. 24.

[7] L. Sweeney, "K-anonymity: a Model For Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5. pp. 557–570, 2002.

[8] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proceedings - IEEE Symposium on Security and Privacy*, 2009, pp. 173–187.

[9] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel, "A practical attack to de-anonymize social network users," in *Proceedings - IEEE Symposium on Security and Privacy*, 2010, pp. 223–238.

[10] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 721–735, 2009.

[11] T. L. Saaty, "Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process," *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, vol. 102, no. 2. pp. 251–318, 2008.

[12] L. A. F. Park and K. Ramamohanarao, "An analysis of latent semantic term self-correlation," *ACM Transactions on Information Systems*, vol. 27, no. 2. pp. 1–35, 2009.

[13] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR*, 1999, pp. 50–57.

[14] Y. Wang and R. K. Nepali, "Privacy Measurement for Social network actor model," in *5th ASE/IEEE International conference on Information Privacy, Security, Risk and Trust*, 2013.

[15] SemEval, "SemEval Portal," 2015. [Online]. Available: http://www.siglex.org/.

[16] H. Nissenbaum, "A Contextual Approach to Privacy Online," *Daedalus*, vol. 140, no. 4. pp. 32–48, 2011.