Privacy Preserving Data Publishing

Author:
Fayola PETERS

Advisor:
Dr. Tim MENZIES

February 2, 2011

1 Introduction

Data mining is the enemy of privacy preservation. In the wrong hands data mining can lead to an individual's identity being stolen, loans and health insurance coverage can be denied and an individual can become a victim to financial fraud. However, in the right hands, terrorists can be identified, scientists find treatments which work better than others, and doctors are aided in treatment and diagnosis decisions. So, to minimize the problems that can be caused by attackers (those seeking to gain confidential knowledge about individuals), the new and active fields of Privacy Preserving Data Publishing (PPDP) and Privacy Preserving Data Mining (PPDM) have been developed.

Researchers in PPDP and PPDM have disclosed their concerns of an attackers' negative use of such a helpful tool. For instance, Wang [WMJH08] wrote:

> Recent development in privacy-preserving data mining has proposed many efficient and practical techniques for hiding sensitive patterns or information from been discovered by data mining algorithms.

Zhang [ZZ07] also noted that although successful in many applications, data mining poses special concerns for private data. Other researchers such [HS10, LS09, ABGP08, Han06], have gone further, showing how easily an attacker could uncover an individuals private/confidential data using readily available data mining tools. So, in order to take advantage of the good which can result from data mining while protecting the privacy of individuals, researchers have come up with many privacy protection techniques in the areas of PPDP and PPDM.

PPDP covers methods and tools used to disguise raw data for publishing, while PPDM covers methods and tools used to limit the amount of extra information gained by an attacker after performing some data mining task on published data. In spite of their relatively short existence, there are a plethora of published privacy protecting techniques designed to address a number of different scenarios. Therefore, to narrow the scope of this paper, we will focus on PPDP and the simple scenario of disguising one data-set for publication.

To explore this, the paper answers the following questions:

- What is PPDP?
- Why is PPDP important?
- What is considered a breach of privacy?
- What are the methods used to avoid privacy breaches?
- How are privacy preserving algorithms evaluated?

2 What is PPDP?

Similarly, researchers agree that PPDP involves the use of techniques to disguise the micro-data records which contain information about specific individuals, while delivering a useful data-set for analysis. This is indicated clearly by Fung [FWCY10], who noted that PPDP is a collection of methods and tools for publishing data in a hostile environment so that the published data remains practically useful while individual privacy is preserved. Along the same lines, Domingo et. al. [DFGN10] states:

> Statistical disclosure control (also known as privacy-preserving data mining) of microdata is about releasing data-sets containing the answers of individual respondents protected in such a way that: (i) the respondents corresponding to the released records cannot be re-identified; (ii) the released data stay analytically useful.

Also, according to LeFevre [LDR08], protecting individual privacy is an important problem in microdata distribution and publishing. Anonymization algorithms typically aim to satisfy certain privacy definitions with minimal impact on the quality of the resulting data.

3 Why is PPDP important?

The importance of PPDP can be answered with a question: *What is data mining without data?*

It is established that data mining is beneficial to many domains such as medical, security, credit fraud and many others. Unfortunately, if the data-sets used in these data mining tasks include a lot of sensitive individual data, the results can lead to privacy breaches. A popular example is the re-identification of William Weld, the former Governor of the state of Massachusetts. Values in quasi identifiers (QIDs) (discussed in Section 5) were linked to his record in a published medical database. More recently, when AOL released the query logs of its users for the purpose of research, Ms. Thelma Arnold was re-identified by the examination of query terms [FWCY10].

Such breaches lowers the confidence of data providers in publishing disguised data with the expectation of no breaches. In the previous example AOL had no choice but to quickly remove the data-set. So a drawback is that data holders are less likely to publish data for research and the benefits of data mining are not realized in these domains.

4 What is considered a breach of privacy?

In order to realize privacy breaches, one needs to define what constitutes a privacy breach for a particular data-set. There are different levels of privacy. Some levels are determined by individuals in the data-set or by the creators of a privacy policy. An optimal result of a privacy model is defined by Dalenius [Dal77] where he states that access to published data should not enable the attacker to learn anything extra about any target victim, compared to no access to the database, even with the presence of any attacker's background knowledge obtained from other sources [FWCY10]. Unfortunately, achieving this ideal is impossible, and so, solutions have focused on trade-offs.

Fung [FWCY10], considers two categories of privacy models, 1) considers that a privacy threat occurs when an attacker is able to link a record owner to a record in a published data table to a sensitive attribute in a published data table, to the published data table itself. These are specified as, record linkage, attribute linkage and table linkage respectively. 2) The published data should provide the attacker with little additional information beyond the background knowledge.

Brickell [BS08], defines a privacy model called sensitive attribute disclosure which occurs when the attacker or adversary learns information about an individual's sensitive attributes. In other words, it captures the gain in the adversaries knowledge due to his observations of the disguised data-set. Also "Microdata privacy can be understood as prevention of membership disclosure" where the attacker should not learn whether a particular individual is included in the database.

In 2007, Wang [WPJ07] put forward other definitions of privacy. The article explains:

> There have been two types of privacy concerning data mining. The first type of privacy, called output privacy, is that the data is minimally altered so that the mining result will not disclose certain privacy. The second type of privacy, called input privacy, is that the data is manipulated so that the mining result is not affected or minimally affected.

For the scenario used in this work, we assume a privacy definition of *high*, where an attacker is unsuccessful at getting more information from the sanitized data-set. The methods used are described in the following section.

5 What are the methods used to avoid privacy breaches?

The idea of disguising a data-set is known as anonymization. This is performed on the original data-set to "satisfy a specified privacy requirement" [FWCY10] resulting in a modified data-set being published. There are five general categories for anonymization, 1) generalization, 2) suppression, 3) anatomization, 4) permutation and 5) perturbation. Most methods and tools created for preserving privacy fall into one or more of these categories.

Before we expand on the above categories, we introduce some information about the structure of the original data-sets:

D(Explicit-Identifier, Quasi-Identifier, Sensitive-Attributes, Non-Sensitive-Attributes), where Explicit-Identifier is a set of attributes, such as name and social security number, containing information that explicitly identifies record owners; Quasi-Identifier (QID) is a set of attributes that could potentially identify record owners; Sensitive-Attributes consists of sensitive person-specific information such as disease, salary, and disability status; and Non-Sensitive-Attributes contains all attributes that do not fall into the previous three categories [FWCY10].

5.1 Generalization and suppression

According to [FWCY10], generalization or suppression hides some details in QID. If attribute values are numerical, generalization resembles discretization in that exact values are replaced by an interval that covers the exact values. With categorical values, specific values are replaced by general ones, for instance, date of birth, becomes month of birth, or $professional$ replaces $engineer$ and $lawyer$. Suppression replaces some values with a special value. This could be looked upon as $blocking$ where special values are replace by a question mark (?) [VBF$^+$04].

$k-anonimity$ is one of the methods which makes each record in the table be indistinguishable with k-1 other records by suppression or generalization [PS10]. There are limitations with k-anonymity including the fact that it does not hide whether a given individual is in the database, or that it does not protect against attacks based on background knowledge [BS08]. To overcome these problems, researchers have proposed many variations of k-anonymity [PS10, MRM10, ZLW09].

5.2 Anatomization and permutation

Anatomization and permutation both accomplish a similar task, that is, the de-association of the relationship between the QID and the sensitive attributes. However, anatomization does it by releasing "the data on QID and the data on the sensitive attribute in two separate tables..." with "one common attribute, *GroupID*" [FWCY10]. On the other hand, permutation de-associates the relationship between a QID and a numerical sensitive attribute. This is done by partitioning a set of data records into groups and shuffling the sensitive values within each group [ZZ07].

5.3 Perturbation

A precise definition of perturbation is put forward by [FWCY10]:

> The general idea is to replace the original data values with some synthetic data values, so that the statistical information computed from the perturbed data does not differ significantly from the statistical information computed from the original data.

It is important to note that the perturbed data records do not correspond to real world record owners. Also, methods used for perturbation include, additive noise, data swapping and synthetic data generation.

6 How are privacy preserving algorithms evaluated?

Three evaluation methods are outlined in [TEM09]. These measures are 1) Information loss measures, 2) Disclosure risk measures and 3) Scores. According to [TEM09], information loss measures are designed to establish in which extent published data is still valid for carrying out the experiments planned on the original data. They take into account the similarity between the original data set and the protected one, as well as the differences between the results that would be obtained with the original data set and the results that would be obtained from the disguised dataset. [TEM09] further explains that disclosure risk measures are used to evaluate the extent in which the protected data ensures privacy and that *scores*, is a summary both information loss and disclosure risk, that is, when these two measures are commensurate, it is possible just to combine them using the average.

7 Conclusions and Future Work

We have presented an overview of PPDP with regards to a single data-set. The importance of reliable methods in PPDP is made clear, in that, if data providers have no confidence in the privacy methods proposed, they will not publish data for the purpose of research. In that vein, these data providers should be provided with not only, methods and tools to disguise their data before being published, but also evaluation measures to help them decided how effective the chosen tool is and whether or not they accept the level of privacy protection offered by the technique.

References

[ABGP08] Maurizio Atzori, Francesco Bonchi, Fosca Giannotti, and Dino Pedreschi. Anonymity preserving pattern discovery. *VLDB JOURNAL*, 17(4):703–727, JUL 2008.

[BS08] Justin Brickell and Vitaly Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 70–78, New York, NY, USA, 2008. ACM.

[Dal77] T Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 1977.

[DFGN10] Josep Domingo-Ferrer and Ursula Gonzalez-Nicolas. Hybrid microdata using microaggregation. *INFORMATION SCIENCES*, 180(15):2834–2844, AUG 1 2010.

[FWCY10] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM COMPUTING SURVEYS*, 42(4), JUN 2010.

[GT09] Aristides Gionis and Tamir Tassa. k-Anonymization with Minimal Loss of Information. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 21(2):206–219, FEB 2009.

[Han06] DJ Hand. Protection or privacy? Data mining and personal data. In Ng, WK and Kitsuregawa, M and Li, J and Chang, K, editor, *ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS*, volume 3918 of *LECTURE NOTES IN ARTIFICIAL INTELLIGENCE*, pages 1–10. 2006.

[HS10] Ayca Azgin Hintogdlu and Yuecel Saygin. Suppressing microdata to prevent classification based inference. *VLDB JOURNAL*, 19(3):385–410, JUN 2010.

[LDR08] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Workload-aware anonymization techniques for large-scale datasets. *ACM TRANSACTIONS ON DATABASE SYSTEMS*, 33(3), AUG 2008.

[LS09] Xiao-Bai Li and Sumit Sarkar. Against Classification Attacks: A Decision Tree Pruning Approach to Privacy Protection in Data Mining. *OPERATIONS RESEARCH*, 57(6):1496–1509, NOV-DEC 2009.

[MRM10] Nissim Matatov, Lior Rokach, and Oded Maimon. Privacy-preserving data mining: A feature set partitioning approach. *INFORMATION SCIENCES*, 180(14):2696–2720, JUL 15 2010.

[PS10] Hyoungmin Park and Kyuseok Shim. Approximate algorithms with generalizing attribute values for k-anonymity. *INFORMATION SYSTEMS*, 35(8):933–955, DEC 2010.

[TEM09] Vicenc Torra, Yasunori Endo, and Sadaaki Miyamoto. ON THE COMPARISON OF SOME FUZZY CLUSTERING METHODS FOR PRIVACY PRESERVING DATA MINING: TOWARDS THE DEVELOPMENT OF SPECIFIC INFORMATION LOSS MEASURES. *KYBERNETIKA*, 45(3):548–560, 2009.

[VBF$^+$04] VS Verykios, E Bertino, IN Fovin, LP Provenza, Y Saygin, and Y Theodoridis. State-of-the-art in privacy preserving data mining. *SIGMOD RECORD*, 33(1):50–57, MAR 2004.

[WMJH08] Shyue-Liang Wang, Rajeev Maskey, Ayat Jafari, and Tzung-Pei Hong. Efficient sanitization of informative association rules. *EXPERT SYSTEMS WITH APPLICATIONS*, 35(1-2):442–450, JUL-AUG 2008.

[WPJ07] Shyue-Liang Wang, Bhavesh Parikh, and Ayat Jafarl. Hiding informative association rule sets. *EXPERT SYSTEMS WITH APPLICATIONS*, 33(2):316–323, AUG 2007.

[XZHW06] Shuting Xu, Jun Zhang, Dianwei Han, and Jie Wang. Singular value decomposition based data distortion strategy for privacy protection. *KNOWLEDGE AND INFORMATION SYSTEMS*, 10(3):383–397, OCT 2006.

[ZLW09] Dan Zhu, Xiao-Bai Li, and Shuning Wu. Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining. *DECISION SUPPORT SYSTEMS*, 48(1, Sp. Iss. SI):133–140, DEC 2009.

[ZZ07] Nan Zhang and Wei Zhao. Privacy-preserving data mining systems. *COMPUTER*, 40(4):52+, APR 2007.