

# Towards Privacy Preserving Data Publishing\*

Xiaoxun Sun<sup>†</sup>

Department of Mathematics & Computing  
University of Southern Queensland  
Toowoomba, Queensland, Australia  
Email: sunx@usq.edu.au

## Abstract

High quality and useful knowledge is to be found in the integrated data from various organizations, and the discovered knowledge is essential for building intelligent systems such as business analysis and health surveillance. However, concern about breaching privacy is a major obstacle of this process. This project aims to develop new efficient and effective techniques for privacy protection in data sharing and data mining by combining techniques in data mining and security research. We focus primarily on notions of anonymity that are defined with respect to individual identity, or with respect to the value of a sensitive attribute. Our goal is to propose a variety of techniques to anonymize original data sets, while preserving the utility of the input data. We adopt extensive evaluations to indicate that it is possible to distribute high-quality data that respects several meaningful notions of privacy. Further, it is possible to do this efficiently for large transactional data sets. The developed cutting edge techniques will advance and facilitate data mining within many organizations and businesses and lead to the better utilization of information.

## 1 Introduction

Anonymity is an important concept for privacy. Anonymity refers to a state that one's identity is completely hidden, and anonymity is oftentimes used as a synonym for privacy. With respect to information privacy, anonymity can embed privacy protection in data itself; for example, no one can tell to whom a data record is related (referred to as *identity privacy*) or no one can learn about a particular property of individuals (referred to as *attribute privacy*) from observing an anonymous dataset. Data anonymity is particularly important in public databases such as census data or health records collected by government agencies, as the information, if linked to individuals, could be highly sensitive. Another situation where data anonymity can be useful is, for example, when an organization wants to allow third parties (e.g., external consulting firms or partner organizations) to access its customer data (or even release such data). Note that in such a case, it cannot be guaranteed that the privacy policy of the data will be always respected. Thus, the organization must make sure that data access cannot violate customers' privacy. The traditional approach of releasing the data tables without breaching the privacy of individuals in the table is to de-identify records by removing the identifying fields such as name, address, and social security number. However, joining this de-identified table with a

publicly available database (like the voters database) on attributes like race, age, and zip code (usually called quasi-identifier) can be used to identify individuals. Famous attacks include de-anonymization of a Massachusetts hospital discharge database by joining it with a public voter database [21] and privacy breaches caused by AOL search data [6].

In order to protect privacy, Sweeney [21, 15, 20] proposed the  $k$ -anonymity model, where some of the quasi-identifier fields are suppressed or generalized so that, for each record in the modified table, there are at least  $k - 1$  other records in the modified table that are identical to it along the quasi-identifier attributes.

Assume that a hospital wants to release patients' medical records in Table 1, referred to as the microdata. Attribute Disease is sensitive, that is, the hospital must ensure that no adversary can correctly infer the disease of any patient with significant confidence. Age, Sex and Zipcode are the quasi-identifier (QI) attributes, because they may be utilized in combination to reveal the identity of an individual, leading to privacy breach. For instance, consider an adversary who has the personal details (i.e., age 22 and Zipcode 4352) of a patient and knows that he has been hospitalized before. In Table 1, since only tuple 3 matches that patient's QI-values, the adversary asserts that the patient contracted Depression. To avoid this problem, generalization divides tuples into QI-groups and transforms their QI-values into less specific forms, so that tuples in the same QI-group cannot be distinguished by their QI-values. Table 2 is a generalized version of Table 1 (e.g., the age 22 and Zipcode 4352 of tuple 3 have been replaced with interval [22-25] and 43\*\*, respectively). Here, generalization produces two QI-groups, including tuples 1-3 and 4-6, respectively. As a result, even if an adversary has the exact QI values of the patient, s/he still does not know which tuple in the first QI-group belongs to him.

## 2 Related Work

In recent years, numerous algorithms have been proposed for implementing  $k$ -anonymity via generalization and suppression. Samarati [15] presents an algorithm that exploits a binary search on the domain generalization hierarchy to find minimal  $k$ -anonymous table. Sun *et al.* [16] recently improve Samarati's algorithm by integrating the hash-based technique. Bayardo and Agrawal [3] present an optimal algorithm that starts from a fully generalized table and specializes the dataset in a minimal  $k$ -anonymous table, exploiting ad hoc pruning techniques. LeFevre *et al.* [8] describe an algorithm that uses a bottom-up technique and a priori computation. Fung *et al.* [5] present a top-down heuristic to make a table to be released  $k$ -anonymous. As to theoretical results, Meyerson and Williams [12] and Aggarwal *et al.* [1]

\*This research is supported by Australian Research Council (ARC) grant DP0774450 and DP0663414.

<sup>†</sup>PhD student enrolled in May, 2007, under the supervision of Dr. Hua Wang, Dr. Ashley Plank and A/Prof. Jiuyong Li.

Gender	Age	Zip Code	Diseases
Male	25	4370	Hypertension
Male	25	4370	Hypertension
Male	22	4352	Depression
Female	28	4373	Chest Pain
Female	28	4373	Obesity
Female	34	4350	Flu

Table 1: The microdata

Gender	Age	Zip Code	Diseases
Male	[22-25]	43**	Hypertension
Male	[22-25]	43**	Hypertension
Male	[22-25]	43**	Depression
Female	[28-34]	43**	Chest Pain
Female	[28-34]	43**	Obesity
Female	[28-34]	43**	Flu

Table 2: A 3-anonymous table

prove that optimal  $k$ -anonymity is NP-hard (based on the number of cells and number of attributes that are generalized and suppressed) and describe approximation algorithms for optimal  $k$ -anonymity. Sun *et al.* [17] prove that  $k$ -anonymity problem is also NP-hard even in the restricted cases, which could imply the results in [1, 12] as well. While focusing on identity disclosure,  $k$ -anonymity model fails to protect attribute disclosure [7]. Several models such as  $p$ -sensitive  $k$ -anonymity [22, 23],  $l$ -diversity [11],  $(\alpha, k)$ -anonymity [26] and  $t$ -closeness [10] are proposed in the literature in order to deal with the problem of  $k$ -anonymity. Although these models can achieve privacy properties to some extent, they are not enough for privacy protection. Limitations of these models are stressed in Section 3.1.

A key difficulty of data anonymization comes from the fact that data utility (i.e., data quality) and data privacy are conflicting goals. Intuitively, data privacy can be enhanced by hiding more data values, but it decreases data utility; on the other hand, revealing more data values increases data utility, but it may decrease data privacy. Thus, we need to devise solutions that best address both the utility and the privacy of data.

Publishing high dimensional data is part of daily operations in commercial activities and public services. A classic example of high dimensional data is transaction databases. Examples of transactions are web queries, click streams, emails, market baskets, and medical notes. Such data often contain rich information and are excellent sources for data mining. Narayanan and Shmatikov show in their recent work [13] that an attacker only needs a little bit information of an individual to identify the anonymized movie rating transaction of the individual in the data set. They re-identify Netflix movie ratings using the Internet Movie Database (IMDb) as a source of auxiliary information and successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Such breach occurs when an attacker only needs a little bit information of an individual to re-identify the anonymized rating transaction of the individual in the data set. Existing research on privacy-preserving data publishing focuses on relational data and the objective is to enforce privacy-preserving paradigms (e.g.,  $k$ -anonymity,  $l$ -diversity, etc) while minimizing the information loss incurred in the anonymizing process. However, methods developed on low dimensional relational data are very inefficient on high dimensional and sparse transactional data.

### 3 Existing Problems and Potential Topics

In this section, we provide some potential research topics regarding privacy preserving data publishing.

#### 3.1 Limitations of Current Privacy Principles

As stated before, since  $k$ -anonymity model is not enough to protect sensitive information, several models such as  $p$ -sensitive  $k$ -anonymity [22],  $l$ -diversity

[11],  $(\alpha, k)$ -anonymity [26] and  $t$ -closeness [10] have been proposed.

*Limitation of  $p$ -sensitive  $k$ -anonymity:* The purpose of  $p$ -sensitive  $k$ -anonymity is to protect against attribute disclosure by requiring that there be at least  $p$  different values for each sensitive attribute within the records sharing a combination of quasi-identifier. This approach has the limitation of implicitly assuming that each sensitive attribute takes values uniformly over its domain; that is, that the frequencies of the various values of a confidential attribute are similar. When this is not the case, achieving the required level of privacy may cause a huge data utility loss.

*Limitation of  $l$ -diversity:* The  $l$ -diversity model protects against sensitive attribute disclosure by considering the distribution of the attributes. The approach requires  $l$  “well-represented”<sup>1</sup> values in each combination of quasi-identifiers. This may be difficult to achieve and, like  $p$ -sensitive  $k$ -anonymity, may result in a large data utility loss. Further, as previously identified,  $l$ -diversity is insufficient to prevent similarity attack.

*Limitation of  $t$ -closeness:* The  $t$ -closeness model protects against sensitive attributes disclosure by defining semantic distance among sensitive attributes. The approach requires the distance between the distribution of the sensitive attribute in the group and the distribution of the attribute in the whole data set to be no more than a threshold  $t$ . Whereas Li *et al.* [10] elaborate on several ways to check  $t$ -closeness, no computational procedure to enforce this property is given. If such a procedure was available, it would greatly damage the utility of data because enforcing  $t$ -closeness destroys the correlations between quasi-identifier attributes and sensitive attributes.

### 3.2 Challenges

Although existing privacy principles can well protect both identity disclosure and attribute disclosure, they all suffer in balancing data privacy and utility. If emphasis is given to privacy, loss of information may result in some applications; while if emphasis is given to usefulness of the data, privacy breaching may occur. The great challenge is how to enhance current privacy preserving paradigms while balancing data privacy and utility. Some preliminary work has been successfully done. In [18], we propose a new privacy protection model called  $(p^+, \alpha)$ -sensitive  $k$ -anonymity, where sensitive attributes are first partitioned into categories by their sensitivity, and then the categories that sensitive attributes belong to are published. Our experimental results show that this model could well protect attribute disclosure while maintaining data utility. Further work on this direction includes extending the  $\alpha$  metric to enhance other privacy paradigms to preserve of proximity privacy. Comprehensive experimental studies should be carried out to verify the superiority to previous privacy protection models in terms of data utility and efficiency.

<sup>1</sup>The interpretation of the term “well-represented” can be found in [11].

### 3.3 Anonymizing Transactional Data

Recent identification of a web searcher from published AOL query logs [6] and de-anonymization methodology applied to the Netflix Prize dataset [13] involve high dimensional transactional databases. Unlike structured records, each transaction contains a small subset of items drawn from a large universe characterized by high dimensionality and sparsity, which makes the existing privacy preserving approaches unsuitable. There are few previous work that consider the privacy of large rating data. In collaboration with MovieLens recommendation service, Frankowski et al. [4] correlate public mentions of movies in the MovieLens discussion forum with the users' movie rating histories in the internal MovieLens dataset. Recent study reveals a new type of de-anonymization for transactional data [13]. Movie ranking data supposedly anonymized is de-identified by linking un-anonymized data from other source. Such transactional data is unstructured, does not have fixed quasi-identifier as relational data, and is characterized by high dimensionality and sparseness.

#### 3.3.1 Challenges

There are several challenges in this topic. First, there is no natural quasi-identifier that is assumed by all previous works, and how to define the decent privacy principle is the first most important issue. Different from quasi-identifier in relational database, the privacy requirement of transactional database should adopt the largeness and sparsity characteristics of transactional database. Second, a direct application of quasi-identifier-based anonymization renders the inefficiency and data useless. Methods developed on low dimensional relational data will be very inefficient on high dimensional, and sparse transactional data set. Faced with this new area, we have finished some first-step work on it. We identified the privacy issues in Netflix Database and present two models to protect privacy, namely,  $(k, \epsilon)$ -anonymity and  $(\epsilon, l)$ -dispense models, one based on protecting identity disclosure, the other on sensitive attribute disclosure. To our best knowledge, there is no model for such data yet. In order to efficiently anonymize the dataset, we adopt the slicing technique to develop our algorithms by utilizing the sparseness of the high dimensional data. Part of the results is submitted to SDM'09 [19]. Further research targets on extending anonymization method to general transactional data.

## 4 Methodology and Significance

In this section, we briefly discuss the approaches that we adopt in dealing with the previous potential research problems and the significance of our research.

### 4.1 Developing $K$ -Anonymity Algorithms

(a) *Hash-based Technique*:  $k$ -anonymity is a technique that prevents "linking" attacks by generalizing and/or suppressing portions of the released microdata so that no individual can be uniquely distinguished from a group of size  $k$ . We investigate a practical model of  $k$ -anonymity, called full-domain generalization. We examine the issue of computing minimal  $k$ -anonymous table based on the definition of minimality described by Samarati. We introduce a novel hash-based technique previously used in mining associate rules and present an efficient hash-based algorithm to find the minimal  $k$ -anonymous table, which improves the previous binary search algorithm first proposed by Samarati.

(b) *Restricted  $K$ -Anonymity*: We introduce two new variants of the  $k$ -anonymity problem, namely, the *Restricted  $k$ -anonymity problem* and *Restricted  $k$ -anonymity problem on attribute* and discuss the connection between the *Restricted  $k$ -anonymity* and the general  $k$ -anonymity problems which stresses the significance of investigating this new class of anonymity problem. We prove that both the *Restricted  $k$ -anonymity problem* and the *Restricted  $k$ -anonymity problem on attribute* are NP-hard. The theoretical results for *Restricted  $k$ -anonymity problem* also provide an alternative NP-hardness proof of general  $k$ -anonymity problem, which implies the main results obtained in [1, 2, 12]. Through a graphical representation of the microdata table, we develop a polynomial time algorithm for the *Restricted 2-anonymity problem*.

### 4.2 Enhancing $K$ -Anonymity Model

$k$ -anonymity alone is not enough to protect privacy in data. Our task is to build effective models that are stronger than the  $k$ -anonymity model and that protect both sensitive facts and private knowledge in data. We introduce a  $(p^+, \alpha)$ -sensitive  $k$ -anonymity model. The  $(p^+, \alpha)$ -sensitive  $k$ -anonymity model requires that in each combination of quasi-identifiers, there are at least  $p$  different sensitive values and the total weight in each combination of quasi-identifiers is at least  $\alpha$ . The motivation for this model is the fact that although  $k$ -anonymity is effective in protecting identity disclosure, to some extent, it fails to protect sensitive attribute disclosure. In the  $(p^+, \alpha)$ -sensitive  $k$ -anonymity model [18], we introduce an ordinal distance system to evaluate the degree that the sensitive attribute contributes to the database.

### 4.3 Anonymizing Large Rating Data

Recent study shows that privacy is at risk when a public movie ranking data set is linked to an anonymous ranking data set. We study a general problem to protect privacy of personal record in a public ranking data set from being attacked by semantic linking using information from other sources like social networks. We propose two novel privacy principles called  $(k, \epsilon)$ -anonymity and  $(\epsilon, l)$ -dispense. Intuitively, the principles require that every transaction has at least  $(k - 1)$   $\epsilon$ -proximate neighbors and for all transactions in its  $\epsilon$ -proximate neighborhood, the average deviation of ratings for each sensitive issue is greater than  $l$ . Noticing the sparse property, we adopt a novel slicing technique to develop anonymization algorithms. We show experimentally, using real-life datasets, that our method is efficient.

### 4.4 Significance

This project addresses an important problem in the information age and aims to develop fundamental and crucial techniques to solve the problem. Data sharing in data mining environments is one of the core technologies needed in many intelligent systems with applications in areas such as health surveillance, business analysis, fraud detection, etc. Improved techniques for privacy protection and data sharing advance and facilitate data mining within many organizations and businesses. However, breaching privacy is a major risk for data sharing and a major obstacle for wide applications of data mining techniques, such as identifying public health problem outbreaks and fraud detection. Individual privacy concerns often prevent the sharing of data among organizations. We plan to investigate this important problem from

a new angle and have some novel ideas to approach the problem. We expect the project to produce some novel techniques for privacy preserving data sharing and data mining.

## 5 Research Outcomes

Five conference papers and one journal paper have been published or accepted related to my research work.

- (1). X. Sun, M. Li, H. Wang and A. Plank. An efficient hash-based algorithm for minimal  $k$ -anonymity problem. *31st Australasian Computer Science Conference (ACSC 2008)*, Wollongong, NSW, Australia. CRPIT 74, pp: 101-107.
- (2). X. Sun, H. Wang and J. Li. On the complexity of restricted  $k$ -anonymity problem. *10th Asia Pacific Web Conference (APWeb 2008)*, LNCS 4976, pp: 287-296, Shenyang, China.
- (3). X. Sun, H. Wang, J. Li, T. M. Truta and P. Li.  $(p^+, \alpha)$ -sensitive  $k$ -anonymity: a new enhanced privacy protection model. *In 8th IEEE International Conference on Computer and Information Technology (IEEE-CIT 2008)*, 8-11 July 2008, Sydney, Australia. pp:59-64.
- (4). X. Sun, H. Wang and J. Li. Priority driven  $k$ -anonymisation for privacy protection. *to appear in 8th Australasian Data Mining Conference (AusDM 2008)*, 27-28 November, 2008, Adelaide, Australia. CRPIT 87.
- (5). X. Sun, H. Wang and J. Li.  $L$ -diversity based updating technique for large time-evolving micro-data. *to appear in 21st Australasian Joint Conference on Artificial Intelligence (AusAI 2008)*, LNAI 5360, pp: 461-470, 3-5 December 2008, Auckland, New Zealand.
- (6). X. Sun, H. Wang, J. Li and T. M. Truta. Enhanced  $P$ -sensitive  $k$ -anonymity models for privacy preserving data publishing. *to appear in Transactions on Data Privacy, 2008*

Above is a brief introduction of my intended work during the PhD study and we will target at top level conference and high impact journal paper publication. We appreciate any expert suggestion concerning privacy preserving data mining and knowledge discovery.

## References

- [1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas and A. Zhu. Anonymizing tables. *In Proc. of the 10th International Conference on Database Theory (ICDT05)*, pp. 246-258, Edinburgh, Scotland.
- [2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas and A. Zhu. Approximation algorithms for  $k$ -anonymity. *Journal of Privacy Technology*, paper number 20051120001.
- [3] R. Bayardo and R. Agrawal. Data privacy through optimal  $k$ -anonymity. *In Proceedings of the 21st International Conference on Data Engineering (ICDE)*, 2005.
- [4] D. Frankowski, D. Cosley, S. Sen, L. G. Terveen and J. Riedl. You are what you say: privacy risks of public mentions. *SIGIR 2006*: 565-572.
- [5] B. Fung, K. Wang, P. Yu. Top-down specialization for information and privacy preservation. *In Proc. of the 21st International Conference on Data Engineering (ICDE05)*, Tokyo, Japan.
- [6] S. Hansell. AOL removes search data on vast group of web users. *New York Times*, Aug 8 2006.
- [7] D. Lambert. Measure of disclosure risk and harm. *Journal of Official Statistics*, vol 9, 1993, pp. 313-331.
- [8] K. LeFevre, D. DeWitt and R. Ramakrishnan. Incognito: Efficient Full-Domain  $k$ -Anonymity. *In ACM SIGMOD International Conference on Management of Data*, June 2005.
- [9] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional  $k$ -anonymity. *In Proceedings of the 22nd International Conference on Data Engineering*, 2006
- [10] N. Li, T. Li and S. Venkatasubramanian.  $t$ -Closeness: Privacy Beyond  $k$ -Anonymity and  $l$ -Diversity. *ICDE 2007*: 106-115
- [11] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian.  $l$ -Diversity: Privacy beyond  $k$ -anonymity. *In ICDE*, 2006.
- [12] A. Meyerson and R. Williams. On the complexity of optimal  $k$ -anonymity. *In Proc. of the 23rd ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems*, pp. 223-228, Paris, France, 2004.
- [13] A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. *to appear in IEEE Security & Privacy 2008*.
- [14] D. J. Newman, S. Hettich, C. L. Blake and C. J. Merz. UCI Repository of Machine Learning Databases, available at [www.ics.uci.edu/~mllearn/MLRepository.html](http://www.ics.uci.edu/~mllearn/MLRepository.html), University of California, Irvine, 1998.
- [15] P. Samarati. Protecting respondents' identities in micro-data release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010-1027. 2001
- [16] X. Sun, M. Li, H. Wang and A. Plank. An efficient hash-based algorithm for minimal  $k$ -anonymity problem. *31st Australasian Computer Science Conference (ACSC 2008)*, Wollongong, NSW, Australia. CRPIT 74, pp: 101-107.
- [17] X. Sun, H. Wang and J. Li. On the complexity of restricted  $k$ -anonymity problem. *10th Asia Pacific Web Conference (APWEB2008)*, LNCS 4976, pp: 287-296, Shenyang, China.
- [18] X. Sun, H. Wang, J. Li, T. M. Truta and P. Li.  $(p^+, \alpha)$ -sensitive  $k$ -anonymity: a new enhanced privacy protection model. *In: 8th IEEE International Conference on Computer and Information Technology*, 8-11 July 2008, Sydney, Australia. pp:59-64
- [19] X. Sun, J. Li, H. Wang and J. Pei. Anonymizing large rating data. *Submitted to SIAM International Conference on Data Mining (SDM09)*.
- [20] L. Sweeney.: Achieving  $k$ -Anonymity Privacy Protection Using Generalization and Suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based System*, 10(5) pp. 571-588, 2002.
- [21] L. Sweeney.  $k$ -Anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty Fuzziness Knowledge-based Systems*, 10(5), pp 557-570, 2002
- [22] T. M. Traian and V. Bindu, Privacy Protection:  $p$ -Sensitive  $k$ -Anonymity Property *International Workshop of Privacy Data Management (PDM2006)*, *In Conjunction with 22th International Conference of Data Engineering (ICDE)*, Atlanta, 2006.
- [23] T. M. Truta, A. Campan and P. Meyer. Generating Micro-data with  $P$ -sensitive  $k$ -anonymity Property. *SDM 2007*: 124-141
- [24] K. Wang, P. S. Yu, and S. Chakraborty.: Bottom-up Generalization: A Data Mining Solution to Privacy Protection. *The fourth IEEE International Conference on Data Mining (ICDM2004)* 249-256.
- [25] W. E. Winkler. Advanced Methods for Record Linkage, *Proceedings of the Section on Survey Research Methods, American Statistical Society*, 467-472, 1994
- [26] R. Wong, J. Li, A. Fu and K. Wang.  $(\alpha, k)$ -anonymity: an enhanced  $k$ -anonymity model for privacy preserving data publishing. *KDD 2006*: 754-759.