2009 International Symposium on Computing, Communication, and Control (ISCCC 2009)
Proc .of CSIT vol.1 (2011) © (2011) IACSIT Press, Singapore

Survey of Anonymity Techniques for Privacy Preserving

Luo Yongcheng[1, 2], Le Jiajin[2] and Wang Jian [2+]

[1] *Library, Donghua University, Shanghai, 201620, China*
[2] *College of Information Science and Technology, Donghua University, Shanghai, 201620, China*

Abstract. Protecting data privacy is an important problem in microdata distribution. Anonymity techniques typically aim to protect individual privacy, with minimal impact on the quality of the released data. Recently, a few of models are introduced to ensure the privacy protecting and/or to reduce the information loss as much as possible. That is, they further improve the flexibility of the anonymous strategy to make it more close to reality, and then to meet the diverse needs of the people. Various proposals and algorithms have been designed for them at the same time. In this paper we provide an overview of anonymity techniques for privacy preserving. We discuss the anonymity models, the major implementation ways and the strategies of anonymity algorithms, and analyze their advantage and disadvantage. Then we give a simple review of the work accomplished. Finally, we conclude further research directions of anonymity techniques by analyzing the existing work.

Keywords: anonymity techniques, privacy preserving, anonymity models, algorithm.

1. Introduction

With the development of data analysis and processing technique, the privacy disclosure problem about individual or enterprise is inevitably exposed when releasing or sharing data, then give the birth to the research issue on privacy preserving. To protect privacy against re-identifying individuals by joining multiple public data sources, after a technique of privacy preserving called k-anonymity [1] was proposed by Samarati and Sweeney in 1998, the anonymity techniques became one of the most important research issue on privacy preserving. Anonymity techniques typically aim to protect individual privacy, with minimal impact on the quality of the resulting data.

We provide here an overview of anonymity techniques for privacy preserving. The rest of this paper is arranged as follows: Section 2 introduces some strategies and models of anonymity techniques; Section 3 describes the implementation ways and algorithms of the existing anonymity techniques; Conclusion and prospect are drawn in the last section.

2. Anonymity Models

K-anonymization techniques have been the focus of intense research in the last few years. In order to ensure anonymization of data while at the same time minimizing the information loss resulting from data modifications, several extending models are proposed, which are discussed as follows.

2.1. K-anonymity

K-anonymity is one of the most classic models, which technique that prevents joining 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*. In the *k*-anonymous tables, a data set is *k*-anonymous ($k \geqslant 1$) if

[+] Corresponding author. Tel.: + 86-021-67792224; fax: +86-021-67792221.
E-mail address: lycluo@dhu.edu.cn

each record in the data set is indistinguishable from at least (*k*-1) other records within the same data set. The larger the value of *k*, the better the privacy is protected. K-anonymity can ensure that individuals cannot be uniquely identified by linking attacks.

Let *T* (i.e. TABLE Ⅰ) is a relation storing private information about a set of individuals. The attributes in *T* are classified in four categories: an identifier (AI), a sensitive attribute (SA), d quasi-identifier attributes (QI) and other unimportant attributes.

For example, we have a raw medical data set as in TABLE Ⅰ. Attributes sex, age and postcode form the quasi-identifier. Two unique patient records 1 and 2 may be re-identified easily since their combinations of sex, age and postcode are unique. The table is generalized as a 2-anonymous table as in TABLE Ⅱ. This table makes the two patients less likely to be re-identified.

TABLE I. RAW MEDICAL DATA SET

AI	**QI**			**SA**
Name	***Sex***	***Age***	***Postcode***	***Illness***
Bill	M	20	13000	Flu
Ken	M	24	13500	HIV
Linda	F	26	16500	Fever
Mary	F	28	16400	HIV

TABLE II. A 2-A NONYMOS DATA SET OF TABLE I

AI	**QI**			**SA**
Name	***Sex***	***Age***	***Postcode***	***Illness***
Bill	M	[20,24]	13*00	Flu
Ken	M	[20,24]	13*00	HIV
Linda	F	[26,28]	16*00	Fever
Mary	F	[26,28]	16*00	HIV

However, while *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure by the homogeneous attack and the background knowledge attack.

2.2. Extending models

Since *k*-anonymity does not provide sufficient protection against attribute disclosure. The paper in [2] proposes the model of *l*-diversity. The notion of *l*-diversity attempts to solve this problem by requiring that each equivalence class has at least *l* well-represented values for each sensitive attribute. The technology of *l*-diversity has some advantages than *k*-anonymity. Because *k*-anonymity dataset permits strong attacks due to lack of diversity in the sensitive attributes. In this model, an equivalence class is said to have *l*-diversity if there are at least *l* well-represented values for the sensitive attribute.

Because there are semantic relationships among the attribute values, and different values have very different levels of sensitivity. An extending model called *t*-closeness is proposed in [3], which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. That is, a table is said to have *t*-closeness if all equivalence classes have *t*-closeness.

The paper in [4] extends the k-anonymity model to the (α, *k*)-anonymity model to limit the confidence of the implications from the quasi-identifier to a sensitive value (attribute) to within α in order to protect the sensitive information from being inferred by strong implications. After anonymization, in any equivalence class, the frequency (in fraction) of a sensitive value is no more than α.

The paper in [5] proposes such a k^m-anonymization model for transactional databases. Assuming that the maximum knowledge of an adversary is at most m items in a specific transaction, it wants to prevent him from distinguishing the transaction from a set of k published transactions in the database.

LeFevre et al. in [6] propose the notion of multidimensional k-anonymity [7] where data generalization is over multi-dimension at a time, and [8] extend multidimensional generalization to anonymize data for a specific task such as classification.

Recently, m-invariance is introduced by Xiaokui Xiao and Yufei Tao in [9] in order to effectively limit the risk of privacy disclosure in re-publication. The paper in [10] proposes a generalization technique called *HD*-composition to offer protection on serial publishing with permanent sensitive values. It involves two major roles, holder and decoy. Decoys are responsible for protecting permanent sensitive value holder which is a dynamic setting.

According k-anonymity does not take into account personal anonymity requirements, personalized anonymity model is also introduced in [11], The core of the model is the concept of personalized anonymity, i.e., a person can specify the degree of privacy protection for her/his sensitive values.

2.3. Conclusion

In conclusion, from the k-anonymity to the *l*-diversity, and then *t*-closeness, anonymity techniques becomes better security in the every development of the anonymous strategies. K-anonymity can resist the links attack, *l*-diversity can resist against the homogeneous attack [2], *t*-closeness will be able to minimize the loss information which the attacker can obtain from the data sheet. At the same time, *m*-invariance, personalized generalization and weighted anonymous strategy further improve the flexibility of the strategy to make it more close to reality, and then to meet the diverse needs of the people.

However, the existing methods based on the anonymity techniques are inadequate because they cannot guarantee privacy protection in all cases, and often incur unnecessary information loss by performing excessive generalization or suppression. Although personalized generalization improves the flexibility of the anonymous strategy, it is not the truth in the real life that personalized generalization thinks of what the definition of sensitive property is exactly same in a table the same as other strategies. Therefore, how to solve the definition of different sensitive information problems in the same table, namely, researching the anonymous strategy of dynamically assigning the sensitive information, will be very meaningful [12].

As the anonymity strategies become more and more perfect, the ability of the anonymity techniques protecting privacy information safety, gradually improves. Thus in this field there are still many questions need further study.

3. Implementation Ways and Algorithms

In order to prevent any privacy breach and cause the minimum information loss at the same time, some effective implementation ways and algorithms of anonymity techniques have been proposed and review here.

3.1. The implementation ways

According to the different implementation methods of the current anonymity techniques, they can be divided into two major categories. The first category is compression and suppression. Generalization involves replacing (or recoding) a value with a less specific but semantically consistent value. It makes its meaning becomes more widely. For example, the original ZIP codes {201620, 201625} can be generalized to 20162*, thereby stripping the rightmost digit and semantically indicating a larger geographical area. Suppression involves not releasing a value at all. That is, this method directly deleting some property values or records from the data tables.

The other category is clustering and partitioning. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The papers [13] research the anonymity techniques based on clustering. The partitioning methodology as a special kind of clustering is proposed in [14]. Their common characteristics are follows the process. First, it generates equivalence group in accordance with the original data table. Then it further locally abstracts in accordance with the records of the equivalent group. That is, the two equivalent property values in the different equivalence groups may be abstract to a different region. However, it is worth mentioning that these ways generally used together with the first category ways in order to achieve the desired effect.

3.2. K-anonymity

Since the introduction of k-anonymity by Samarati and Sweeney in 1998, a few of algorithms [6, 15, 16, 17, 18, 19] have been proposed to obtain k-anonymous released data. For example, Xu et al. [19] propose some greedy methods to achieve k-anonymity with cell generalization to attain the more less information loss. These algorithms can be divided into two categories, according to the constraints imposed on generalization. The first category employs “full-domain generalization” [6], which assumes a hierarchy on each *QI* attribute, and requires that all the partitions in a general domain should be at the

same level of the hierarchy. That is, this algorithm maps a given value in a single domain to another one globally.

The second category can be termed "full-subtree recoding" [6] drops the same-level requirement mentioned earlier, since it often leads to unnecessary information loss [11, 17]. In full-subtree recoding, the data space is partitioned into a set of (nonoverlapping) regions, and the anonymization maps all tuples in a region to the same generalized or changed tuple. Following this idea, Iyengar [17] develops a genetic algorithm, whereas greedy algorithms are proposed in [16] and [18], based on top-down and bottom-up generalization, respectively. These approaches, however, do not minimize information loss. Bayardo and Agrawal in [15] remedy the problem with the power-set search strategy. Subsequently, [11] extends it to incorporate customized privacy needs in order to attain personalized privacy preservation.

The generalization hierarchy [14, 15] is generally applied to the various generalization strategies. In [6], a nonoverlapping generalization-hierarchy is first defined for each attribute of quasi-identifier. Then an algorithm tries to find an optimal solution which is allowed by such generalization hierarchies. Note that in these schemes, if a lower level domain needs to be generalized to a higher level domain, all the values in the lower domain are generalized to the higher domain. This restriction could be a significant drawback in that it may lead to relatively high data distortion due to unnecessary generalization. The algorithms in [15, 20], on the other hand, allow values from different domain levels to be combined to represent a generalization. Although this leads to much more flexible generalization, possible generalizations are still limited by the imposed generalization hierarchies.

Besides the generalization hierarchies, FLeFevre et al. in [16] transform the k-anonymity problem into a partitioning problem. Specifically, their approach consists of the following two steps. The first step is to find a partitioning of the dimensional space, where n is the number of attributes in the quasi-identifier, such that each partition contains at least k records. Then the records in each partition are generalized so that they all share the same quasi-identifier value. Otherwise, in the paper [21] proposes an approach that uses the idea of clustering to minimize information loss and thus ensure good data quality. The key idea here is that data records that are naturally similar to each other should be part of the same equivalence class [21].

The previous studies show that the problem of optimal anonymity algorithms is *NP*-hard under various quality models [21, 22]. Thus, in the real work, we should propose the most suitable algorithms to attain the special anonymous tables that meet the needs of the privacy protecting and the quality of analysis.

4. Conclusion and Prospect

In this paper we have surveyed several extending *k*-anonymity for preventing re-identification attacks in microdata release. Besides, we also analyze the merits and shortcomings of these techniques. Obviously, anonymity technique is a promising approach for data publishing without compromising individual privacy or disclosing sensitive information. And at present, anonymity techniques are at the stage of development. Some approaches and models are proposed, however, these techniques need be further researched because of the complexity of the privacy problem. We conclude two research directions of the anonymity techniques by analyzing the existing work in the future.

1) The research of the requirement-oriented anonymity techniques for privacy preservation will become the issue. That is, our attention should be paid to the special requirements of individual.

2) How to improve the efficiency of implementation and ensure the quality of the released data in order to meet the various requirements.

5. References

[1] P. Samarati and L. Sweeney, *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,* In Technical Report SRI-CSL-98-04, 1998.

[2] A.Machanavajjhala, J.Gehrke, et al., *ℓ-diversity: Privacy beyond k-anonymity*, In Proc. of ICDE, Apr.2006.

[3] N. Li, T. Li, and S. Venkatasubramanian, *t-Closeness: Privacy Beyond k-anonymity and l-Diversity*, In Proc. of ICDE, 2007, pp. 106-115.

[4] Wong R C, Li J, Fu A W, et al, *(α, k)-Anonymity：an enhaned k-anonymity model for privacy-preserving data publishing*, Proceedings of the 12th ACM SIGKDD, New York: ACM Press, 2006, pp. 754-759.

[5] Terrovitis, M., Mamoulis, N., and Kalnis, *Privacy preserving Anonymization of Set-valued Data,* In VLDB, 2008, pp. 115-125.

[6] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, *Incognito: Efficient full-domain k-anonymity*, In Proceedings of the ACM SIGMOD International Conference on Management of Data, 2005, pp. 49-60.

[7] Xiaojun Ye, Jin, L, Bin Li, *A Multi-Dimensional K-Anonymity Model for Hierarchical Data*, Electronic Commerce and Security, 2008 International Symposium, Aug. 2008, pp. 327-332.

[8] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan., *Workload-aware anonymization*, In Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, August 2006.

[9] Xiao X, Tao Y, *M-invariance: towards privacy preserving re-publication of dynamic datasets*, In Proc. of SIGMOD, New York: ACM Press, 2007, pp. 689-700.

[10] Yingyi Bu, Ada Wai-Chee Fu,et al, *Privacy Preserving Serial Data Publishing By Role Composition*, In VLDB, 2008, pp. 845-856.

[11] Xiao X, Tao Y, *Personalized privacy preservation*, Proceedings of ACM Conference on management of Data (SIGMOD). New York: ACM Press, 2006, pp. 785-790.

[12] LIU Ming, Xiaojun Ye, *Personalized K-anonymity*, Computer Engineering and Design, Jan.2008, pp. 282-286.

[13] Byun J-W, Kamra A, Bertinoe, et al., *Efficient k-anonymization using clustering techniques*, In DASFAA 2007, LNCS 4443.Berlin: Sp ringer-Verlag, 2007, pp.188-200.

[14] LeFevre K, DeWitt D J, Ramakrishnan R, *Mondrian multidimensional K-anonymity*, In Proc. of the International Conference on Data Engineering(ICDE'06), Atlanta, GA, USA, April.2006, pp. 25-35.

[15] R. Bayardo and R. Agrawal, *Data privacy through optimal kanonymization*,In ICDE, 2005, pp. 217–228.

[16] B. C. M. Fung, K. Wang, and P. S. Yu., *Top-down specialization for information and privacy preservation*. In ICDE, 2005, pp. 205–216.

[17] V. Iyengar., *Transforming data to satisfy privacy constraints*, In SIGKDD, 2002, pp. 279–288.

[18] K. Wang, P. S. Yu, and S. Chakraborty., *Bottom-up generalization: A data mining solution to privacy protection*, In ICDM, 2004, pp. 249–256.

[19] Xu J, Wangw, Pei J, et al., *Utility-based anonymization using local recoding*, In Proc. of the 12th ACM SIGKDD. New York: ACM Press, 2006, pp. 785-790.

[20] V. S. Iyengar., *Transforming data to satisfy privacy constraints*, In ACM Conference on Knowledge Discovery and Data mining, 2002.

[21] Ji-Won Byun, Ashish Kamra, et al., *Efficient k-Anonymization Using Clustering Techniques*, In Internal Conference on DASFAA, Berli: Spring-Verlag, April. 2007, pp. 188-200.

[22] G. Aggarwal, T. Feder, K. Kenthapadi, et al., *Anonymizing tables*, In International Conference on Database Theory, 2005, pp.246–258.