Satisfying Privacy Requirements: One Step before Anonymization

Xiaoxun Sun[1], Hua Wang[1], and Jiuyong Li[2]

[1] Department of Mathematics & Computing
University of Southern Queensland, Australia
{sunx,wang}@usq.edu.au
[2] School of Computer and Information Science
University of South Australia, Australia
jiuyong.li@unisa.edu.au

Abstract. In this paper, we study a problem of privacy protection in large survey rating data. The rating data usually contains both ratings of sensitive and non-sensitive issues, and the ratings of sensitive issues include personal information. Even when survey participants do not reveal any of their ratings, their survey records are potentially identifiable by using information from other public sources. We propose a new (k, ϵ, l)-anonymity model, in which each record is required to be similar with at least $k-1$ others based on the non-sensitive ratings, where the similarity is controlled by ϵ, and the standard deviation of sensitive ratings is at least l. We study an interesting yet nontrivial satisfaction problem of the (k, ϵ, l)-anonymity, which is to decide whether a survey rating data set satisfies the privacy requirements given by users. We develop a slice technique for the satisfaction problem and the experimental results show that the slicing technique is fast, scalable and much more efficient in terms of execution time than the heuristic pairwise method.

1 Introduction

The problem of privacy-preserving data publishing has received a lot of attention in recent years. Privacy preservation on tabular data has been studied extensively. A major category of privacy attacks on relational data is to re-identify individuals by joining a published table containing sensitive information with some external tables. Most of existing work can be formulated in the following context: several organizations publish detailed data about individuals (e.g. medical records) for research or statistical purposes [14,10,9,13].

Privacy risks of publishing microdata are well-known. Famous attacks include de-anonymisation of the Massachusetts hospital discharge database by joining it with a public voter database [14] and privacy breaches caused by AOL search data [5]. Even if identifiers such as names and social security numbers have been removed, the adversary can use linking [12], homogeneity and background attacks [10] to re-identify individual data records or sensitive information of individuals. To overcome the re-identification attacks, k-anonymity was proposed

M.J. Zaki et al. (Eds.): PAKDD 2010, Part I, LNAI 6118, pp. 181–188, 2010.

[12,14]. Specifically, a data set is said to be k-anonymous ($k \geq 1$) if, on the quasi-identifier attributes, each record is identical with at least $k-1$ other records. The larger the value of k, the better the privacy is protected. Several algorithms are proposed to enforce this principle [2,6,7,8]. Machanavajjhala et al. [10] showed that a k-anonymous table may lack of diversity in the sensitive attributes. To overcome this weakness, they propose the l-diversity [10]. However, even l-diversity is insufficient to prevent attribute disclosure due to the skewness and the similarity attack. To amend this problem, t-closeness [9] was proposed to solve the attribute disclosure vulnerabilities inherent to previous models.

Recently, a new privacy concern has emerged in privacy preservation research: how to protect individuals' privacy in large survey rating data. Though several models and many algorithms have been proposed to preserve privacy in relational data (e.g., k-anonymity [12,14], l-diversity [10], t-closeness [9], etc.), most of the existing studies are incapable of handling rating data, since the survey rating data normally does not have a fixed set of personal identifiable attributes as relational data, and it is characterized by high dimensionality and sparseness. The survey rating data shares the similar format with transactional data. The privacy preserving research of transactional data has recently been acknowledged as an important problem in the data mining literature [3,15,16]. To our best knowledge, there is no current research addressing the issue of how to efficiently determine whether the survey rating data satisfies the privacy requirement.

2 Motivation

On October 2, 2006, Netflix, the world's largest online DVD rental service, announced the $1-million Netflix Prize to improve their movie recommendation service [4]. To aid contestants, Netflix publicly released a data set containing 100,480,507 movie ratings, created by 480,189 Netflix subscribers between December 1999 and December 2005. Narayanan and Shmatikov shown in their recent work [11] that an attacker only needs a little information to identify the anonymized movie rating transaction of the individual. They re-identified 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 political preferences and other potentially sensitive information.

We consider the privacy risk in publishing survey rating data. For example, in a life style survey, ratings to some issues are non-sensitive, such as the likeness of a book. Ratings to some issues are sensitive, such as the income level. Assume that each survey participant is cautious about his/her privacy and does not reveal his/her ratings. However, it is easy to find his/her preferences on non-sensitive issues from publicly available information sources, such as personal weblog. An attacker can use these preferences to re-identify an individual and consequently find sensitive ratings of a victim. An example is given in the Table 1. In a social network, people make comments on various issues, which are not considered sensitive. Some comments can be summarized as in Table 1(b). We assume that people are aware of their privacy and do not reveal their ratings,

Table 1. (a) A published survey rating data (b) Public comments on some non-sensitive issues of some survey participants

	non-sensitive			sensitive
ID	issue 1	issue 2	issue 3	issue 4
t_1	6	1	*null*	6
t_2	1	6	*null*	1
t_3	2	5	*null*	1
t_4	1	*null*	5	1
t_5	2	*null*	6	5

(a)

	non-sensitive issues		
name	issue 1	issue 2	issue 3
Alice	excellent	so bad	-
Bob	awful	top	-
Jack	bad	-	good

(b)

either non-sensitive or sensitive ones. However, individuals in the supposedly anonymized survey rating data are potentially identifiable based on their public comments from other sources. For example, Alice is at risk of being identified, since the attacker knows Alice's preference on issue 1 is 'excellent', by cross-checking Table 1(a) and (b), s/he will deduce that t_1 in Table 1(a) is linked to Alice, the sensitive rating on issue 4 of Alice will be disclosed. This example motivates us the following research question: Given a survey rating data set T with the privacy requirements, how to efficiently determine whether T satisfies the given privacy requirements?

3 (k, ϵ, l)-Anonymity

We assume that a survey rating data set publishes people's ratings on a range of issues. Each survey participant is cautious about his/her privacy and does not reveal his/her ratings. However, an attacker may find a victim's preference (not exact rating scores) by personal familiarity or by reading the victim's comments on some issues from personal weblog. We consider that attackers know preferences of non-sensitive issues of a victim but do not know exact ratings and want to find out the victim's ratings on some sensitive issues.

The auxiliary information of an attacker includes: (i) the knowledge of a victim being in the survey rating data; (ii) preferences of the victims on some non-sensitive issues. The attacker wants to find ratings on sensitive issues of the victim. In practice, knowledge of Types (i) and (ii) can be gleaned from an external database [11]. For example, in the context of Table 1(b), an external database may be the IMDb. By examining the anonymized data in Table 1(a), the adversary can identify a small number of candidate groups that contain the record of the victim. It will be the unfortunate scenario where there is only one record in the candidate group. For example, since t_1 is unique in Table 1(a), Alice is at risk of being identified. If the candidate group contains not only the victim but other records, an adversary may use this group to infer the sensitive value of the victim. For example, although it is difficult to identify whether t_2 or t_3 in Table 1(a) belongs to Bob, since both records have the same sensitive value, Bob's private information is identified.

In order to avoid such attack, we propose a (k, ϵ, l)-anonymity model. The first step is to require that in the released data, every transaction should be *similar* with at least $k-1$ other records based on the non-sensitive ratings so that no survey participants are identifiable. For example, t_1 in Table 1(a) is unique, and based on the preference of Alice in Table 1(b), her sensitive issues can be re-identified. Jack's sensitive issues, on the other hand, is much safer. Since t_4 and t_5 in Table 1(a) form a similar group based on their non-sensitive rating. Second is to prevent the sensitive rating from being inferred by requiring the sensitive ratings in a similar group be diverse. For instance, although t_2 and t_3 in Table 1(a) form a similar group, their sensitive ratings are identical. Therefore, an attacker can immediately infer Bob's preference on the sensitive issue without identifying which transaction belongs to Bob. In contrast, Jack's preference on the sensitive issue is much safer than both Alice and Bob.

Given a rating data set T, each transaction contains a set of numbers indicate the ratings on some issues. Let $(o_1, \cdots, o_p, s_1, \cdots, s_q)$ be a transaction, $o_i \in \{1 : r, null\}$, $i = 1, 2, \cdots, p$ and $s_j \in \{1 : r, null\}$, $j = 1, 2, \cdots, q$, where r is the maximum rating and *null* indicates that a survey participant did not rate. $o_1, \cdots, o_p$ stand for non-sensitive ratings and $s_1, \cdots, s_q$ denote sensitive ratings. Let $T_A = \{o_{A_1}, \cdots, o_{A_p}, s_{A_1}, \cdots, s_{A_q}\}$, $T_B = \{o_{B_1}, \cdots, o_{B_p}, s_{B_1}, \cdots, s_{B_q}\}$ be the ratings for participants A and B, then dissimilarity of non-sensitive ratings ($Dis(o_{A_i}, o_{B_i})$ and sensitive ratings ($Dis(s_{A_i}, s_{B_i})$) between T_A and T_B is defined as follows:

$$Dis(o_{A_i}, o_{B_i}) = \begin{cases} |o_{A_i} - o_{B_i}| & o_{A_i}, o_{B_i} \in \{1 : r\} \\ 0 & o_{A_i} = o_{B_i} = null \\ r & \text{otherwise} \end{cases} \quad (1)$$

$$Dis(s_{A_i}, s_{B_i}) = \begin{cases} |s_{A_i} - s_{B_i}| & s_{A_i}, s_{B_i} \in \{1 : r\} \\ r & s_{A_i} = s_{B_i} = null \\ r & \text{otherwise} \end{cases} \quad (2)$$

Definition 1 (ϵ-proximate). *Given a survey rating data set T with a small positive number ϵ, two transactions $T_A = \{o_{A_1}, \cdots, o_{A_p}, s_{A_1}, \cdots, s_{A_q}\} \in T$ and $T_B = \{o_{B_1}, \cdots, o_{B_p}, s_{B_1}, \cdots, s_{B_q}\} \in T$. T_A and T_B are ϵ-proximate, if $\forall 1 \leq i \leq p$, $Dis(o_{A_i}, o_{B_i}) \leq \epsilon$. The set of transactions T is ϵ-proximate, if every two transactions in T are ϵ-proximate.*

If two transactions are ϵ-proximate, the dissimilarity between their non-sensitive ratings is bound by ϵ.

Definition 2 ((k, ϵ)-anonymity). *A survey rating data set T is said to be (k, ϵ)-anonymous if and only if every transaction is ϵ-proximate with at least $k-1$ other transactions. The transaction t with all the other transactions that are ϵ-proximate with t in T form a (k, ϵ)-anonymous group.*

Although (k, ϵ)-anonymity can protect identity, it fails to protect sensitive information. For example, in Table 1(a), t_2 and t_3 are in a $(2, 1)$-anonymous group,

but they have the same rating on the sensitive issue, thus Bob's private information is breaching. This example reflects the shortcoming of the (k, ϵ)-anonymity. To mitigate this limitation, sufficient diversity of the sensitive values in each (k, ϵ)-anonymous group should be allowed.

For a sensitive issue s, let the vector of ratings of the group be $(s_1, \cdots, s_g)$, where $s_i \in \{1 : r, null\}$. The mean of the ratings is $\bar{s} = \frac{1}{Q}\sum_{i=1}^{g} s_i$, where Q is the number of non-*null* values, and $s_i \pm null = s_i$. The standard deviation of the rating is then defined as $SD(s) = \sqrt{\frac{1}{g}\sum_{i=1}^{g}(s_i - \bar{s})^2}$.

Definition 3 ((k, ϵ, l)-anonymity). *A survey rating data set is said to be (k, ϵ, l)-anonymous if and only if the standard deviation of sensitive ratings is at least l in each (k, ϵ)-anonymous group.*

4 The Algorithm

In this section, we formulate the satisfaction problem and develop a slicing technique to determine the following *Satisfaction Problem.*

Satisfaction problem: Given a survey rating data T and k, ϵ, l, the satisfaction problem of (k, ϵ, l)-anonymity is to decide whether T satisfies k, ϵ, l requirements.

The satisfaction problem is to determine whether the user's given privacy requirement is satisfied by the given data set. If the data set has already met the requirements, it is not necessary to make any modifications before publishing. As follows, we propose a novel slice technique to solve the satisfaction problem.

Algorithm 1: $Slicing(\epsilon, T, t_0)(C)$

```
1  Can ← {t_0}; S ← ∅
2  / * To slice out the ε-proximate of t_0 * /
3  for j ← 1 to n
4  do if |t_j − t_0| ≤ ε
5        then Cand ← Cand ∪ {t_j}
6             S ← S ∪ {j}
7  / * To trim the ε-proximate of t_0 * /
8  PCand ← Cand
9  for i ← 1 to |S|
10 do for j ← 1 to |S|
11    do if |t_S(i) − t_S(j)| > ε
12          then PCand ← PCand \ {t_S(i)}
13 return PCand
```

We illustrate the slicing technique using an example in 3-D space. Given $t = (t_1, t_2, t_3) \in T$, our goal is to slice out a set of transactions T $(t \in T)$ that are ϵ-proximate. Our approach is first to find the ϵ-proximate of t, which is the set of transactions that lie inside a cube C_t of side 2ϵ centered at t. Since ϵ is typically

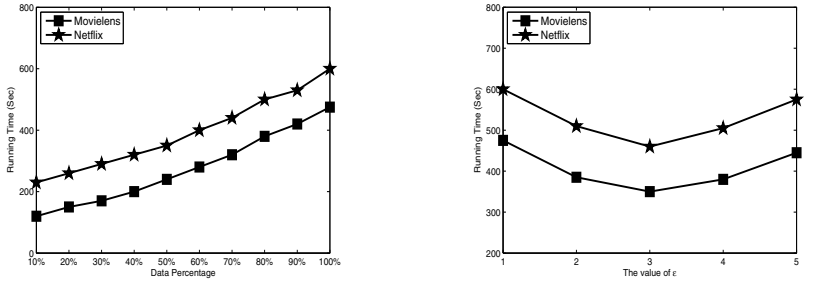
small, the number of points inside the cube is also small. The ϵ-proximate of C'_t can be found by an exhaustive comparison within the ϵ-proximate of t. If there are no transactions inside the cube C_t, we know the ϵ-proximate of t is empty, so as the ϵ-proximate of the set C'_t. The transactions within the cube can be found as follows. First we find the transactions that are sandwiched between a pair of parallel planes X_1, X_2 and add them to a *candidate set*. The planes are perpendicular to the first axis of coordinate frame and are located on either side of the transaction t at a distance of ϵ. Next, we trim the candidate set by disregarding transactions that are not also sandwiched between the parallel pair of Y_1 and Y_2, that are perpendicular to X_1 and X_2. This procedure is repeated for Z_1 and Z_2 at the end of which, the candidate set contains only transactions within the cube of size 2ϵ centered at t. $Slicing(\epsilon, T, t_0)$ (Algorithm 1) describes how to find the ϵ-proximate of the set C_{t_0} with $t_0 \in C_{t_0}$.

5 Experiments

Our experimentation deploys the MovieLens data downloadable at `http://www.grouplens.org/taxonomy/term/14`, which was made available by the GroupLens Research Project at the University of Minnesota. In the data set, a user is considered as an object while a movie is regarded as an attribute and many entries are empty since a user only rated a small number of movies.

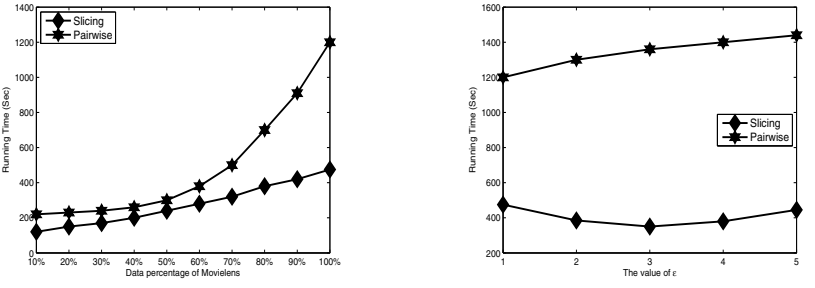
Data used for Fig. 1(a) is generated by re-sampling the Movielens data set while varying the percentage of data from 10% to 100%. We evaluate the running time for the (k, ϵ, l)-anonymity model with default setting $k = 20, \epsilon = 1, l = 2$. The execution time for (k, ϵ, l)-anonymity is increasing with the increased data percentage. This is because as the percentage of data increases, the computation cost increases too, since the overhead is increased with the more dimensions. Next, we evaluate how the parameters affect the cost of computing. In these experiments, we use the whole MovieLens data and evaluate by varying ϵ. With $k = 20, l = 2$, Fig. 1(b) shows the computational cost as a function of ϵ, in determining (k, ϵ, l)-anonymity. At the initial stage, when ϵ is small, more computation efforts are put into finding ϵ-proximate of the transactions, but less used in exhaustive search for ϵ-proximate of the set, and this explains the initial decent of overall cost. As ϵ grows, the searching time for ϵ-proximate is reduced, but the number of transactions in the ϵ-proximate is increased, which results in huge exhaustive search effort and this causes the eventual cost increase.

In addition to the scalability, we experimented the comparison between the slicing algorithm (Slicing) and the heuristic pairwise algorithm (Pairwise), which works by computing all the pairwise distances to construct the dissimilarity matrix and identify the violation of privacy requirements. We implemented both algorithms and studied the impact of the execution time on the data percentage and ϵ. Fig. 2(a) describe the trend of the algorithms by varying the percentage of the data set. From the graph, the slicing algorithm is far more efficient than the heuristic pairwise algorithm especially when the volume of the data becomes larger. This is because, when the dimension of the data increases, the disadvantage of the heuristic pairwise algorithm, which is to compute all the dissimilarity



(a) (b)

Fig. 1. Running time comparison on Movielens data set vs. (a) data percentage varies; (b) ϵ varies



(a) (b)

Fig. 2. Running time comparison of Slicing and Pairwise methods (a) data percentage varies; (b) ϵ varies

distance, dominates the execution time. On the other hand, the smarter grouping technique used in the slicing process makes less computation cost for the slicing algorithm. The similar trend is shown in Fig. 2(b) by varying ϵ.

6 Conclusion

We have studied the problem of protecting individuals' sensitive ratings in the large survey rating data. We proposed a novel (k, ϵ, l)-anonymity model and studied the satisfaction problem. A novel slicing technique was proposed to solve the satisfaction problem by searching closest neighbors in large, sparse and high dimensional survey rating data. The experimental results confirm the slicing technique is fast and scalable in practical.

Acknowledgement

Thanks for the reviewers' valuable comments. The research is supported by Australian Research Council (ARC) discovery grants DP0774450 and DP0663414.

References

1. Backstrom, L., Dwork, C., Kleinberg, J.: Wherefore Art Thou R3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography. In: WWW 2007, pp. 181–190 (2007)
2. Fung, B., Wang, K., Yu, P.: Top-down specialization for information and privacy preservation. In: ICDE 2005, pp. 205–216 (2005)
3. Ghinita, G., Tao, Y., Kalnis, P.: On the Anonymisation of Sparse High-Dimensional Data. In: ICDE 2008, pp. 715–724 (2008)
4. Hafner, K.: If you liked the movie, a Netflix contest may reward you handsomely. New York Times (2006)
5. Hansell, S.: AOL removes search data on vast group of web users. New York Times (2006)
6. Le Fevre, K., De Witt, D., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: SIGMOD 2005, pp. 49–60 (2005)
7. Le Fevre, K., De Witt, D., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: ICDE 2006, pp. 25–26 (2006)
8. Li, J., Tao, Y., Xiao, X.: Preservation of Proximity Privacy in Publishing Numerical Sensitive Data. In: SIGMOD 2008, pp. 473–486 (2008)
9. Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: Privacy Beyond k-anonymity and l-diversity. In: ICDE 2007, pp. 106–115 (2007)
10. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-Diversity: Privacy beyond k-anonymity. In: ICDE 2006, pp. 24–25 (2006)
11. Narayanan, A., Shmatikov, V.: Robust De-anonymisation of Large Sparse Datasets. In: IEEE Security & Privacy, 111–125 (2008)
12. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
13. Sun, X., Wang, H., Li, J.: Injecting purposes and trust into data anonymization. In: CIKM 2009, pp. 1541–1544 (2009)
14. Sweeney, L.: k-Anonymity: A Model for Protecting Privacy. International Journal on Uncertainty Fuzziness Knowledge-based Systems 10(5), 557–570 (2002)
15. Xu, Y., Wang, K., Fu, A., Yu, P.S.: Anonymizing Transaction Databases for Publication. In: KDD 2008, pp. 767–775 (2008)
16. Xu, Y., Fung, B., Wang, K., Fu, A., Pei, J.: Publishing Sensitive Transactions for Itemset Utility. In: ICDM 2008, pp. 1109–1114 (2008)