2010 Fourth International Conference on Network and System Security

Towards Identify Anonymization in Large Survey Rating Data

Xiaoxun Sun
Department of Mathematics & Computing
University of Southern Queensland, Australia
sunx@usq.edu.au

Hua Wang
Department of Mathematics & Computing
University of Southern Queensland, Australia
wang@usq.edu.au

***Abstract*—We study the challenge of identity protection in the large public survey rating data. Even though the survey participants do not reveal any of their ratings, their survey records are potentially identifiable by using information from other public sources. None of the existing anonymisation principles (e.g., k-anonymity, l-diversity, etc.) can effectively prevent such breaches in large survey rating data sets. In this paper, we tackle the problem by defining the (k, ϵ)-anonymity principle. The principle requires for each transaction t in the given survey rating data T, at least $(k-1)$ other transactions in T must have ratings similar with t, where the similarity is controlled by ϵ. We propose a greedy approach to anonymize survey rating data and apply the method to two real-life data sets to demonstrate their efficiency and practical utility.**

I. Introduction

The problem of privacy-preserving data publishing has received a lot of attention in recent years [9], [15]. Privacy preservation on relational data has been studied extensively. A major type of privacy attack on relational data includes re-identifying individuals by joining a published data set containing sensitive information with the external data sets modeling background knowledge of attackers [14]. Most of the existing work is formulated in contexts of several organizations, such as hospitals, publishing detailed data (also called microdata) about individuals (e.g. medical records) for research or statistical purposes.

Recently, a new privacy concern has emerged in privacy preservation research: how to protect the privacy of individuals in published large survey rating data. For example, movie rating data, supposedly to be anonymized, is de-identified by linking un-anonymized data from another source. On October 2, 2006, Netflix, the world's largest online DVD rental service, announced a $1-million Netflix Prize for improving their movie recommendation service [8]. 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 [15] have shown 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-identified Netflix movie ratings using the Internet Movie Database (IMDb) (http://www.imdb.com/) as a source of auxiliary information and successfully identified the Netflix records of known users, uncovering their political preferences and other potentially sensitive information. In this paper, we will refer to two types of data as "survey rating data" and "relational data".

A. Motivation

Table I(a) is a published survey rating data set containing ratings of survey participants on a range of issues either sensitive or non-sensitive. The higher the rating is, the more preferred the participant is towards the issue. "$null$" means the participant did not rate the issue. Table I(b) contains comments on non-sensitive issues of some survey participants, which might be obtained from public information sources such as personal weblogs or social network.

However, individuals's private ratings in the anonymous survey rating data set are potentially identifiable based on their public comments from other sources [15]. By matching the ratings of non-sensitive issues with publicly available preferences, an adversary can identify a small number of candidate groups that contain the record of the victim. It is unfortunate if there is only one record in the candidate group. For example, Alice is at risk of being identified in Table I(a), since t_1 is unique and could be linked to Alice's comments in Table I(b). This example motivates the following challenge: *How to preserve individual's identity privacy in a large survey rating data set?*

Though several algorithms have been proposed to preserve privacy in relational data, most of them can deal with relational data only [17], [14], [13]. The structure of large survey rating data is different from relational data, since it does not have fixed personal identifiable attributes. The lack of a clear set of personal identifiable attributes makes the anonymisation challenging [23], [7]. In addition, survey rating data contains many attributes, each of which corresponds to the response to a survey question, but not all participants need to rate all issues (or answer all questions), which means a lot of cells in a data set are empty. Hence, previous methods can not be applied to deal with survey rating data and it is much more challenging to devise anonymisation methods for large survey rating data than for relational data.

II. Related Work

Privacy preserving data publishing has received considerable attention in recent years, especially in the context of

978-0-7695-4159-4/10 $26.00 © 2010 IEEE
DOI 10.1109/NSS.2010.11

	non-sensitive			sensitive
ID	issue 1	issue 2	issue 3	issue 4
t_1	6	1	*null*	6
t_2	3	6	*null*	1
t_3	4	5	*null*	4
t_4	2	5	*null*	1
t_5	1	*null*	5	1
t_6	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)

Table I: (a) A published survey rating data set containing ratings of survey participants on both sensitive and non-sensitive issues. b) Public comments on some non-sensitive issues of some participants of the survey.

relational data [12], [11], [1], [16], [14], [13], [19], [20]. All these works assume a given set of attributes QID on which an individual is identified, and anonymize data records on the QID. Aggarwal [1] presents a study on the relationship between the dimensionality of QID and information loss, and concludes that, as the dimensionality of QID increases, information loss increases quickly. Large survey rating data sets present a worst case scenario for existing anonymisation approaches because of the high dimensionality of QID and sparseness of the data sets. To our best knowledge, all existing solutions in the context of k-anonymity [17], [16], l-diversity [14] and t-closeness [13] assume a relational table, which typically has a low dimensional QID. Survey rating data sets, on the other hand, are characterized by sparseness and high dimensionality, which makes the current state-of-art principles incapable handling the anonymisation of large survey rating data sets.

There are few previous works considering the privacy of large rating data. In collaboration with MovieLens recommendation service, Frankowski *et al.* correlated public mentions of movies in the MovieLens discussion forum with the users' movie rating histories in the internal Netflix data set [5]. Recent study reveals a new type of attack on anonymized MovieLens data [15]. The supposedly anonymized movie rating data is re-identified by linking non-anonymized data from other sources. To our best knowledge, no anonymisation models and methods exist for preserving privacy for large survey rating data sets.

Privacy-preservation of transactional data has been acknowledged as an important problem in the data mining literature [3], [4], [21], [7], [23]. The privacy threats caused by publishing data mining results such as frequent item sets and association rules is addressed in [3], [4]. The work in [2], [21] focus on publishing anonymous patterns, where the patterns are mined from the original data, and the resulting set of rules is sanitized to present privacy breaches. In contrast, our work addresses the privacy threats caused by publishing a large survey rating data. Recent work [7], [23] targets anonymisation of transaction data. Our work aims to prevent individual identity disclosure in a large survey rating data set. Our recent work [18] addresses the problem of how to decide whether a survey rating data satisfies the given privacy requirements, we do not discuss techniques for anonymizing survey rating data.

III. (k, ϵ)-ANONYMITY

In this section, we formally define the (k, ϵ)-anonymity model for protecting privacy in large survey rating data.

We assume that survey rating data publishes people's ratings on a range of issues. Some issues are sensitive, such as income level, while some are non-sensitive, such as the opinion of a book, a movie or a kind of food. Each survey participant is cautious about his/her privacy and does not reveal his/her ratings. However, an attacker can use auxiliary information to identify an individual's sensitive ratings in supposedly anonymous survey rating data. The auxiliary information of an attacker includes: (i) knowledge that a victim is in the survey rating data and; (ii) preferences of the victims on some non-sensitive issues. For instance, 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 weblogs or social networks. We assume 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. Our objective is to design an effective model to protect privacy of people's sensitive ratings in published survey rating data.

Given a survey rating data set T, each transaction contains a set of numbers indicating the ratings on some issues. Let $(o_1, o_2, \cdots, o_p, s_1, s_2, \cdots, s_q)$ 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. Each transaction belongs to a survey participant. Let $T_A = \{o_{A_1}, o_{A_2}, \cdots, o_{A_p}, s_{A_1}, s_{A_2}, \cdots, s_{A_q}\}$ be the ratings for a survey participant A and $T_B = \{o_{B_1}, o_{B_2}, \cdots, o_{B_p}, s_{B_1}, s_{B_2}, \cdots, s_{B_q}\}$ be the ratings for a participant B. We define the dissimilarity between two non-sensitive rating scores as follows. $Dis(o_{A_i}, o_{B_i}) = |o_{A_i} - o_{B_i}|$ if $o_{A_i}, o_{B_i} \in \{1 : r\}$; $Dis(o_{A_i}, o_{B_i}) = 0$ if $o_{A_i} = o_{B_i} = null$; otherwise $Dis(o_{A_i}, o_{B_i}) = r$.

Definition 1 (ϵ-proximate). *Given a small positive number ϵ, if for $1 \leq i \leq p$, $Dis(o_{A_i}, o_{B_i}) \leq \epsilon$, transactions T_A and T_B are ϵ-proximate.*

If two transactions are ϵ-proximate, the dissimilarity between their non-sensitive ratings is bound by ϵ. In Table I(a), if $\epsilon = 1$, ratings 5 and 6 may have no difference in interpretation, so t_5 and t_6 are 1-proximate based on their non-sensitive rating.

Definition 2 ((k, ϵ)-anonymity). *A survey rating data set is (k, ϵ)-anonymous if every transaction in the survey rating data set has at least $(k-1)$ ϵ-proximate neighbors.*

The idea behind (k, ϵ)-anonymity is to make each transaction in a survey rating data set similar with at least other $(k-1)$ transactions in order to avoid linking to individual's sensitive ratings. (k, ϵ)-anonymity can well protect identity privacy, since it guarantees that no individual is identifiable with confidence up to a function of ϵ with probability greater than $1/k$. Given a survey rating data set T and the values of k, ϵ, the objective of (k, ϵ)-anonymisation is to modify T to make it satisfy the k, ϵ requirements. In the next section, we discuss identity anonymization in survey rating data.

IV. Identity anonymize in survey rating data

In this section, we describe the anonymization technique of the (k, ϵ)-anonymity model in survey rating data. We first introduce some preliminaries and the metric to quantify the distortion caused by anonymization, and then describe the greedy anonymization algorithm with examples.

A. Preliminaries

Given a survey rating data set T, we define a binary flag matrix $F(T)$ to record if there is a rating or not for each non-sensitive issue. $F(T)_{ij} = 1$ if the ith participant rates the jth issue and $F(T)_{ij} = 0$ otherwise. For instance, the flag matrix associated with the rating data of Table I(a) is

$$\mathbf{F} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad (1)$$

in which each row corresponds to survey participants and each column corresponds to non-sensitive issues. In order to measure the distance between two vectors in the flag matrix, we borrow the concept of Hamming distance [10].

Definition 3 (Hamming Distance). *Hamming distance between two vectors in the flag matrix of equal length is the number of positions for which the corresponding symbols are different. We denote the Hamming distance between two vectors v_1 and v_2 as $H(v_1, v_2)$.*

In other words, Hamming distance measures the minimum number of substitutions required to change one vector into the other, or the number of errors that transformed one vector into the other. For example, if $v_1 = (1, 1, 0)$ and $v_2 = (1, 0, 1)$, then $H(v_1, v_2) = 2$. If the Hamming distance between two vectors is zero, then these two vectors are identical. In order to categorize identical vectors in the flag matrix, we introduce the concept of Hamming group.

Definition 4 (Hamming Group). *Hamming group is the set of vectors in which the Hamming distance between any two vectors of the flag matrix is zero. The maximal Hamming group is a Hamming group that is not a subset of any other Hamming group.*

For example, there are two maximal Hamming groups in the flag matrix (1) made up of vectors $\{(1,1,0),(1,1,0),(1,1,0),(1,1,0)\}$ and $\{(1,0,1),(1,0,1)\}$ and they correspond to groups $\{t_1, t_2, t_3, t_4\}$ and $\{t_5, t_6\}$ of T.

B. Distortion Metrics

In this section, we define a measure to capture the information loss.

Definition 5 (Tuple distortion). *Let $t = (t_1, t_2, \cdots, t_m)$ be a tuple and $t' = (t'_1, t'_2, \cdots, t'_m)$ be an anonymized tuple of t. Then, the distortion of this anonymisation is defined as:*

$$\textit{Distortion}(t, t') = \sum_{i=1}^{m} |t_i - t'_i|$$

For example, if the tuple $t = (5, 6, 0)$ is generalized to $t' = (5, 5, 0)$, then the distortion of this anonymisation is $|5 - 5| + |6 - 5| + |0 - 0| = 1$.

Definition 6 (Total distortion). *Let $T' = (t'_1, t'_2, \cdots, t'_n)$ be the anonymized data set from $T = (t_1, t_2, \cdots, t_n)$. Then, the total distortion of this anonymisation is defined as:*

$$\textit{Distortion}(T, T') = \sum_{i=1}^{n} \textit{Distortion}(t_i, t'_i)$$

For example, let $T = (t_1, t_2, t_3, t_4)$, where $t_1 = (5, 6, 0)$, $t_2 = (2, 5, 0)$, $t_3 = (4, 7, 0)$ and $t_4 = (5, 6, 0)$. Let $T' = (t'_1, t'_2, t'_3, t'_4)$ be anonymization of T, where $t'_1 = (5, 5, 0)$, $t'_2 = (3, 5, 0)$, $t'_3 = (3, 7, 0)$ and $t'_4 = (5, 7, 0)$. Then, the distortion between the two data sets is $1 + 1 + 1 + 1 = 4$.

C. The algorithm

For ease, we first illustrate our approach in the scale of single attribute, and then we extend it to multiple attributes.

Let $t = (t_1, t_2, \cdots, t_n)$ be the ratings of some issue from n survey participants with the privacy requirement ϵ. We assume that some ratings in t are not bounded by ϵ, and our aim is to modify t to make every pair of ratings is bounded by ϵ while minimizing the distortion. The idea of

the approach is as follows. Order all ratings for the issue t, and find the minimum rating Min and maximum rating Max. Find all intervals of the size ϵ between Min and Max. Change the ratings that does not fit in this interval such that the distortion is minimized. In the case of some tuples with the same minimum distortion, randomly pick up one of them as the anonymization. The process is described in **Algorithm 1**.

ALGORITHM 1: $single\_anonymizer(t, \epsilon)$

```
1  Input: an ascended tuple t = (t_1, ..., t_n), and ε
2  Output: t' = (t'_1, ..., t'_n) with minimum distortion
3  / * Computing distortions for all intervals * /
4  for i ← 1 to (t_n − t_1)/ε
5     do for j ← 1 to n
6           do if t_j ∈ (t_i, t_i + ε)
7                 then t'_j ← t_j
8                 else if t_j < t_i
9                       t'_j ← t_i
10                else t'_j ← t_i + ε
11          D(i) ← Distortion(t', t);
12 / * Finding minimum distortion * /
13 k ← 1; D_min ← D(k);
14 for i ← 2 to (t_n − t_1)/ε
15    do if D(i) < D_min
16          then D_min ← D(i);
17               k ← i;
18 / * Retrieving t' with minimum distortion * /
19 for i ← 1 to n
20    do if t_i ∈ (t_k, t_k + ε)
21          then t'_i ← t_i
22          else if t_i < t_k
23                t'_i ← t_k
24          else t'_i ← t_k + ε
25 return t'
```

For example, if $t = (3, 4, 5, 6, 7, 7, 8, 8)$ and $\epsilon = 2$. The Min is 3 and Max is 8. Build all the intervals with the size of 2, which are (3,5), (4,6), (5,7) and (6,8). Following Algorithm 1, the anonymization of t is shown in Table II, in which the vector in bold is the anonymisation we choose.

Intervals	Anonymization	Distortion
(3, 5)	(3, 4, 5, 5, 5, 5, 5, 5)	11
(4, 6)	(4, 4, 5, 6, 6, 6, 6, 6)	7
(5, 7)	**(5, 5, 5, 6, 7, 7, 7, 7)**	**5**
(6, 8)	(6, 6, 6, 6, 7, 7, 8, 8)	6

Table II: Example of the anonymization algorithm

In order to extend the approach to deal with multiple issues, we first find all the maximal Hamming groups in T, there are two cases may happen;

Case 1: The cardinality of each maximal Hamming group is greater than k. For each Hamming group, we apply the algorithm $single\_anonymizer(A, \epsilon)$ on every issue A in T to get the anonymized survey rating data T'.

Case 2: There exists at least one maximal Hamming group containing less than k participant. In this case, we distribute the records to other Hamming groups to make the cardinality of each Hamming group be at least k and follow the procedure of Case 1 to anonymize T.

Let us take Table I(a) as an example with $k = 2, \epsilon = 1$. As discussed in Section IV, there are two maximal Hamming groups $HG_1 = \{t_1, t_2, t_3, t_4\}$ and $HG_2 = \{t_5, t_6\}$. HG_2 has already satisfied the privacy requirement, but HG_1 does not. The anonymization of HG_1 is shown in Table III, in which the vector in bold is the anonymisation we choose.

	Intervals	Anonymization	Distortion
Issue 1	(2,3)	(3,3,3,2)	4
	(3,4)	**(4,3,4,3)**	3
	(4,5)	(5,4,4,4)	4
	(5,6)	(6,5,5,5)	6
	Intervals	Anonymization	Distortion
Issue 2	(1,2)	(1,2,2,2)	10
	(2,3)	(2,3,3,3)	8
	(3,4)	(3,4,4,4)	6
	(4,5)	**(4,5,5,5)**	4
	(5,6)	(5,6,5,5)	4

Table III: Anonymizing HG_1 of Table I(a)

D. Complexity analysis

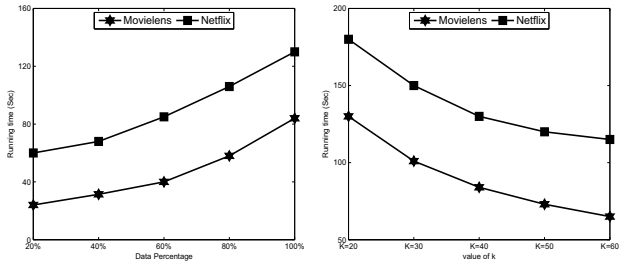
Recall that our objective is to anonymize data consisting of a set of transactions $T = \{t_1, t_2, \cdots, t_n\}$, $|T| = n$. Each transaction $t_i \in T$ contains m issues. The computation cost consists of three parts, which are sorting, finding intervals and computing distortion. The complexity of the sorting is $O(mn\log n)$. During the next phrase of the algorithm, for each attribute, we find the Min and Max and all the possible intervals with size ϵ, which incur the amount of $O(2(n-1))$ overhead, and the cost for comparisons to search the one with least distortion is $O(n)$. So, the total complexity of all attributes in this phrase is $O(mn)$. The last phrase to compare original and anonymous data sets to estimate the distortion has the cost of $O(mn)$. The computational complexity of this alternative approach is $O(mn\log n+mn)$.

V. EXPERIMENTAL STUDY

In this section, we experimentally evaluate the effectiveness and efficiency of the proposed anonymization algorithms.

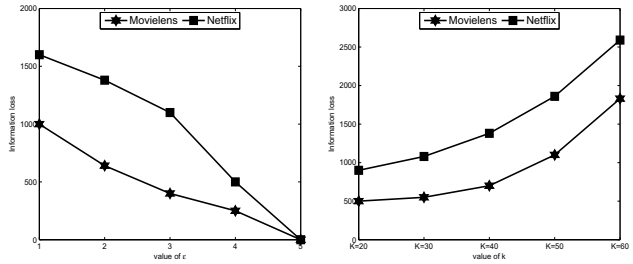
A. Data Sets

Our experimentation deploys two real-world databases. MovieLens (http://www.grouplens.org/taxonomy/term/14 and Netflix data sets (http://www.netflixprize.com/). MovieLens data set contains 100,000 ratings (5-star scale), 943 users and 1682 movies. Netflix data set contain



(a) (b)

Figure 1: Running time on Movielens and Netflix databases vs. (a) Data percentage varies (b) k varies



(a) (b)

Figure 2: Information loss comparison on Movielens and Netflix databases vs. (a) k varies; (b) ϵ varies

over 100,480,507 ratings from 480,189 randomly-chosen, anonymous Netflix customers over 17 thousand movie titles. The ratings are on a scale from 1 to 5 (integral) stars.

B. Efficiency

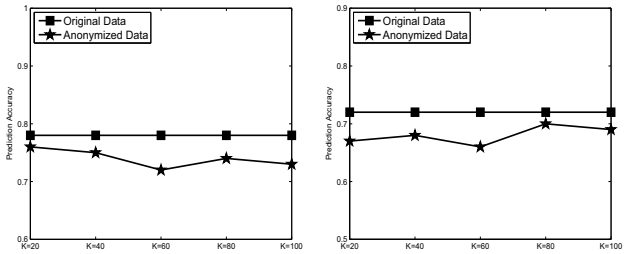
Data used for Figure 1(a) is generated by re-sampling the Movielens and Netflix data sets while varying the percentage from 20% to 100%. We evaluate the running time for the (k, ϵ)-anonymity model with default setting $k = 40, \epsilon = 2$. For both data sets, the execution time for (k, ϵ)-anonymity is increasing with the increasing data percentage. This is because as the percentage of data increases, the computation cost increases too. The result is expected since the overhead is increased with the more dimensions.

Next, we evaluate how the parameters affect the cost of computing. Data set used for this sets of experiments are the whole sets of MovieLens and Netflix data and we evaluate by varying the value of k. Setting $\epsilon = 2$, Figure 1(b) displays the results of running time by varying k from 20 to 60 for both data sets. The cost drops as k grows. This is expected, because fewer search efforts for ϵ-proximate neighborhoods needed for a greater k, allowing our algorithm to terminate earlier.

C. Data Utility

Having verifying the efficiency of our technique, we proceed to test its effectiveness. We measure the utility by the distortion metric defined in Section IV-B. Generally speaking, the more the distortion is, the less useful the anonymized data would be.

We first study the influence of ϵ (i.e., the length of a proximate neighborhood) on data utility. Towards this, we set k to 40. Concerning $(40, \epsilon)$-anonymity, Figure 2(a) plots the information loss on both data sets as a function of ϵ. The anonymization algorithm incurs less distortion as ϵ increases. This is expected, since a smaller ϵ demands stricter privacy preservation, which reduces data utility. When $\epsilon = 5$, there will be no anonymization required, and therefore the information loss reaches 0. Next, we examine the utility of $(k, 2)$-anonymous solution with different k. Figure 2(b) presents the information loss as a function of k. The error grows with k because a larger k demands tighter anonymity control requiring much more data modification.



(a) (b)

Figure 3: Prediction Accuracy: (a) Movielens; (b) Netflix

Figures 3(a) and (b) evaluate the classification and prediction accuracy of the greedy anonymization algorithm. Our evaluation methodology is that we first divide data into training and testing sets, and we apply the anonymization algorithm to the training and testing sets to obtain the anonymized training and testing sets, and finally the classification or regression model is trained by the anonymized training set and tested by anonymized testing set. The Weka implementation [24] of simple Naive Bayes classifier was used for the classification and prediction. Using the Movielens data, Figure 3(a) compares the predictive accuracy of classifier trained on Movielens data produced by the greedy anonymization algorithm. In these experiments, we generated 50 independent training and testing sets, each containing 2000 records, and we fixed $\epsilon = 2$. The results are averaged across these 50 trials. For comparison, we also include the accuracies of classifier trained on the (not anonymized) original data. From the graph, we can see that the average prediction accuracy is around 75%, very close to the original accuracy, which preserves better utility for data mining purposes. Similar results are obtained by using the Netflix rating data in Figure 3(b).

VI. Conclusion and future work

In this paper, we mitigate a privacy threat to a large survey rating data set with a principle called (k, ϵ)-anonymity. We apply the flag matrix to formulate the problem, through which we provide a greedy approach to anonymize rating data. Extensive experiments confirm that our technique produces anonymized data sets that are useful.

This work also initiates several directions for future investigations on our research agenda. First, the (k, ϵ)-anonymity model is targeted at identify protection, it is also important to address the issue of how to prevent attribute disclosures. The privacy principle similar to l-diversity might be considered. Second, it is also interesting to employ dimensionality-reduction techniques for more effective anonymisation.

Acknowledgement

We would like to thank for the reviewers' valuable comments for improving the paper. The research is supported by Australian Research Council (ARC) grant DP0774450 and DP0663414.

References

[1] C. Aggarwal. On k-Anonymity and the curse of dimensionality. VLDB 2005, pp. 901-909.

[2] M. Atzori, F. Bonchi, F. Giannotti and D. Pedreschi. Anonymity preserving pattern discovery. VLDB J. 17(4), pp. 703-727 (2008)

[3] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi. Blocking anonymity threats raised by frequent itemset mining. ICDM 2005, pp.561-564.

[4] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi. k-anonymous patterns. PKDD 2005, pp. 10-21.

[5] D. Frankowski, D. Cosley, S. Sen, L. G. Terveen and J. Riedl. You are what you say: privacy risks of public mentions. SIGIR 2006: pp, 565-572

[6] M. R. Garey and D. S. Johnson. Computers and Intractability: A Guide to the Theory of $\mathcal{NP}$-Completeness. San Francisco. Freeman, 1979

[7] G. Ghinita, Y. Tao and P. Kalnis. On the Anonymisation of Sparse High-Dimensional Data, In Proceedings of International Conference on Data Engineering (ICDE) April 2008, pp. 715-724.

[8] K. Hafner. And if you liked the movie, a Netflix contest may reward you handsomely. New York Times, Oct 2 2006.

[9] S. Hansell. AOL removes search data on vast group of web users. New York Times, Aug 8 2006.

[10] R. W. Hamming. Coding and Information Theory, Englewood Cliffs, NJ, Prentice Hall (1980).

[11] D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In SIGMOD Conference, pp. 217-228, 2006.

[12] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: efficient full-domain k-anonymity. SIGMOD 2005, pp. 49-60.

[13] N. Li, T. Li and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-anonymity and l-diversity. ICDE 2007: pp. 106-115

[14] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-Diversity: Privacy beyond k-anonymity. ICDE 2006, pp. 24.

[15] A. Narayanan and V. Shmatikov. Robust De-anonymisation of Large Sparse Datasets. IEEE Security & Privacy 2008, pp. 111-125.

[16] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6): pp: 1010-1027. 2001.

[17] L. Sweeney. k-Anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty Fuzziness Knowledge-based Systems*, 10(5), pp. 557-570, 2002

[18] X. Sun, H. Wang and J. Li. Satisfying Privacy Requirements: One Step Before Anonymization. *to appear in the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2010)*, 2010.

[19] X. Sun, H. Wang and J. Li. Injecting Purpose and Trust into Data Anonymisation. *in the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, 2009.

[20] X. Sun, L. Sun and H. Wang. Extended k-Anonymity Models Against Sensitive Attribute Disclosure. *to appear in Computer Communications.*, 2010

[21] V. S. Verykios, A. K. Elmagarmid, E. Bertino, E. Dasseni and Y. Saygin. Association Rule Hiding. IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 4, pp. 434-447, April 2004.

[22] R. Wong, J. Li, A. Fu, K. Wang. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. *KDD 2006*: pp, 754-759.

[23] Y. Xu, K. Wang, Ada Wai-Chee Fu and Philip S. Yu. Anonymizing Transaction Databases for Publication. KDD 2008, pp. 767-775.

[24] I. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2nd edition, 2005.