Anonymizing Data with Quasi-Sensitive Attribute Values

Pu Shi
Department of Math & CS
Emory University
Atlanta, GA, USA
mr.pu.shi@gmail.com

Li Xiong*
Department of Math & CS
Emory University
Atlanta, GA, USA
lxiong@mathcs.emory.edu

Benjamin C. M. Fung
CIISE
Concordia University
Montreal, QC, Canada
fung@ciise.concordia.ca

ABSTRACT

We study the problem of anonymizing data with quasi-sensitive attributes. Quasi-sensitive attributes are not sensitive by themselves, but certain values or their combinations may be linked to external knowledge to reveal indirect sensitive information of an individual. We formalize the notion of *l-diversity* and *t-closeness* for quasi-sensitive attributes, which we call *QS l-diversity* and *QS t-closeness*, to prevent indirect sensitive attribute disclosure. We propose a two-phase anonymization algorithm that combines quasi-identifying value generalization and quasi-sensitive value suppression to achieve *QS l-diversity* and *QS t-closeness*.

Categories and Subject Descriptors

H.2.7 [**Database Administration**]: Security, integrity, and protection

General Terms

Algorithms, Design, Experimentation, Security.

1. INTRODUCTION

Privacy-preserving data publishing or data anonymization [2] has received considerable attention in recent years as a promising approach for sharing useful information while preserving data privacy. In a typical non-interactive scenario, a data publisher (e.g. hospital) collects and stores data from record owners (e.g. patients), referred to as microdata, and then releases a sanitized view of the data to a data recipient (e.g. public health researchers) for querying or analysis purposes.

There are two important disclosure risks associated with the released data. *Identity disclosure* occurs if an individual can be identified from released data. *Attribute disclosure* occurs when a sensitive attribute is revealed and can be attributed to an individual. Given a microdata table, several subsets of attributes have been considered in light of the disclosure risks [2]: *identifiers (I)* which explicitly identify record owners and are typically removed from the released data, *quasi-identifiers (QI)* which could be linked with external information to re-identify individual record owners, and *sensitive attributes (S)* which should be protected. Most existing work that aim to prevent attribute disclosure, based on notable principles such as *l*-diversity [5] and *t*-closeness [4], consider explicit sensitive attributes. However, in practice, many attributes, especially set-valued attributes, may not be sensitive by themselves, but certain values or their combinations may be linked to external knowledge to reveal sensitive information of an individual. We refer to this kind of attributes as *quasi-sensitive* attributes.

Age	Gender	State	Symptoms
22	M	GA	Coughing, Headache, Sore Throat
25	M	GA	Coughing, Headache
30	F	TX	Headache, Vomiting
35	F	TX	Headache, Sore Throat

(a) Original database

Age	Gender	State	Coughing	Vomiting	Headache	Sore throat
22	M	GA	Yes	No	Yes	Yes
25	M	GA	Yes	No	Yes	No
30	F	TX	No	Yes	Yes	No
35	F	TX	No	No	Yes	Yes

(b) Original database with set-valued data in binary representations

Disease	Symptoms		
Flu	Coughing	Headache	Sore throat
Hepatitis B	Loss of appetite	Vomiting	Dark urine

(c) External knowledge on symptoms

Age	Gender	State	Symptoms
[20, 30]	M	GA	Coughing, Headache, Sore throat
[20, 30]	M	GA	Coughing, Headache
[30, 40]	F	TX	Headache, Vomiting
[30, 40]	F	TX	Headache, Sore throat

(d) A generalized table

Figure 1: Anonymizing Data with Quasi-Sensitive Attributes

Consider an example shown in Figure 1. Table 1(a) shows an original database containing patients' medical records. *Age, Gender, State* is a set of QI attributes. *Symptoms* is a quasi-sensitive set-valued attribute. Table 1(b) shows an alternative representation of Table 1(a) with Symptoms represented as binary attributes. Table 1(c) shows an external table containing information about diseases and their possible symptoms which are public knowledge. *Symptoms* is not sensitive by its own but its values, when linked to the

*Corresponding author.

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

external table, may reveal the hidden sensitive information of *Disease*.

One may attempt to apply the existing *l-diversity* or *t-closeness* based generalization technique to prevent the potential disclosure of disease. One way is to treat *Symptoms* as a single sensitive attribute. For example, Table 1(d) is a possible generalization of Table 1(a) that satisfies k-anonymity (k=2) and l-diversity (l=2) for *Symptoms*. Consider an adversary who has the QI-values {22, M, GA} of a target individual. Given the external table, s/he can easily infer that the victim has Flu even if the QI group for the victim has diverse symptoms. Alternatively, we can represent *Symptoms* in a binary form as shown in Table 1(b) and treat each individual value as a binary sensitive attribute and apply l-diversity for multiple sensitive attributes [5]. Unfortunately, due to the high dimensionality and sparsity, not only the dataset will be over-generalized but also it is still possible to link the individual to a unique hidden sensitive value.

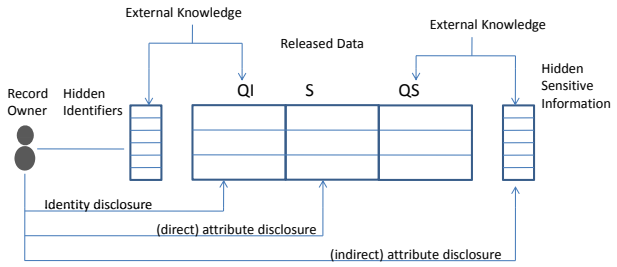


Figure 2: Disclosure Risks with Quasi-Sensitive Attributes

Contributions. This paper provides the first study towards preventing *indirect* sensitive information disclosure due to the *quasi-sensitive* attribute values. Figure 2 illustrates the problem space of related disclosure risks and our contributions. Our first contribution is to formalize the notion of *l-diversity* and *t-closeness* for quasi-sensitive attribute values, which we call *QS l-diversity* and *QS t-closeness*, to prevent indirect attribute disclosure due to external knowledge. Our second contribution is a two-phase algorithm that combines generalization and value suppression to achieve *QS l-diversity* and *QS t-closeness*. In the first phase, generalization on the QI attributes is used to produce an intermediate dataset. In the second phase, we propose a novel and effective QS attribute value suppression algorithm, which uses greedy heuristics but adaptively updates the solution within a time bound. With initial empirical evaluations, we demonstrate that our approach is practical and effective.

2. DEFINITION AND ALGORITHM

2.1 QS l-diversity and QS t-closeness

In this section, we introduce the definitions for quasi-sensitive attributes and our attack scenario and formalize the privacy principles. We use D and D^* to denote the original and released microdata tables respectively.

Definition 1 (Quasi-sensitive attribute values). *A set of attribute values are quasi-sensitive (QS) if they are not sensitive by themselves but can be linked with external data to reveal sensitive information of individuals.*

We aim to prevent indirect attribute disclosure attacks through the QS attributes. In such an attack, we assume an adversary knows: i) the existence of a victim individual tp in D, and ii) the exact QI values of tp, denoted as q. The adversary also has access to an external knowledge table E, similar to that in [1], which associates a set of attribute values (also referred to as *terms* or *items*) to a piece of sensitive information (referred to as *sensitive label*). Formally, each row in E is a pair $(L_i, S_i), i = 1, 2, ..., |E|$, where L_i is a sensitive label and S_i is a corresponding set of quasi-sensitive attribute values.

Definition 2 (Linking to sensitive labels). *A tuple tp in D^* can be linked to a sensitive label L_i if and only if each quasi-sensitive attribute value of this tuple is contained in S_i. All the sensitive labels that can be linked to tp compose $K(tp)$, the sensitive label set of tp. All the sensitive labels that can be linked to all d tuples in a QI group G is $\cup_{i=1}^{d} K(tp_i)$, the sensitive label set of G.*

In our example shown in Figure 1, *Symptoms* is a quasi-sensitive set-valued attribute. The external table of symptoms links the quasi-sensitive attribute values with sensitive information of disease. The first record is linked and uniquely linked to the sensitive label *Flu*. Consider an adversary who has the QI-values {22, M, GA} of a target individual. Given a possible anonymized table in Table 1(d) and the external table in Table 1(c), s/he can easily infer that the victim has Flu.

Below we analyze the prior and posterior beliefs of linking an individual to a sensitive label and define quasi-sensitive l-diversity and t-closeness to bound the change of beliefs.

Prior belief. The attacker's prior belief, denoted by $\alpha_{(q,L)}$, is the probability that a targeted individual tp with QI-vector value q is linked to a label L in the external table before the data release. The attacker's prior belief about the sensitive label results only from his background knowledge. Given the external knowledge table E, s/he could estimate each label's frequencies in the microdata by counting the labels' accumulated frequencies in the mapped knowledge sets.

Posterior belief. After observing the released dataset D^*, the attacker's belief changes. The posterior belief, denoted by $\beta_{(q,L)}$, measures the probability that QI-vector value q (contained in QI-group G in D^*) is linked to a sensitive label L in external table E. We have:

$$\beta_{(q,L)} = P(G \rightarrow L) = \sum_{tp_i \in G} (\frac{1}{|G|} * P(tp_i \rightarrow L))$$

Here $P(tp_i \rightarrow L)$ is the probability that the tuple tp_i in G is linked to L. Obviously, $\sum_{L \in K(tp_i)} (P(tp_i \rightarrow L) = 1$, but the specific value of $P(tp_i \rightarrow L)$ may vary. For simplicity, we assume

$$P(tp_i \rightarrow L) = \begin{cases} 1/|K(tp_i)|, if L \in K(tp_i) \\ 0, otherwise \end{cases}$$

We now define principles *QS l-diversity* and *QS t-closeness* based on the l-diversity and t-closeness principles respectively to bound the difference of the prior and posterior beliefs. For the *l-diversity* definition, since *(c,l)-diversity* is the most powerful and well-adopted version, we define *QS l-diversity* based on *(c,l)-diversity*.

Definition 3 (*QS (c,l)-diversity*). *Suppose we have the values of $\beta_{(q,L)}$ for each of the L in $\cup_{i=1}^{d} K(tp_i)$, and they are p_1, p_2, ..., $p_{|\cup_{i=1}^{d} K(tp_i)|}$ in decreasing order respectively (i.e. $p_1 \geq p_2 \geq ... \geq p_{|\cup_{i=1}^{d} K(tp_i)|}$). The group G is said to satisfy* QS (c,l)-diversity *if and only if $p_1 \leq c * (p_l + p_{l+1} + ... + p_{|\cup_{i=1}^{d} K(tp_i)|})$. A table D^* satisfies QS (c,l)-diversity if every group satisfies QS (c,l)-diversity.*

Definition 4 (*QS t-closeness*). *The group G with QI-value q satisfies* QS t-closeness *if and only if the distance $d(\beta_{(q,*)}, \alpha_{(q,*)})$ between the posterior beliefs $\beta_{(q,L)}$ and the prior beliefs $\alpha_{(q,L)}$ is no more than a threshold t. A table D^* satisfies QS t-closeness if every group satisfies QS t-closeness.*

For simplicity and computational efficiency, we will define the distance as:

$$d(\beta_{(q,*)}, \alpha_{(q,*)}) = (\sum_{\forall L} (\beta_{(q,L)} - \alpha_{(q,L)})^2)^{1/2}$$

2.2 Algorithm

Two-phase anonymization. To achieve the above privacy requirements, we propose an anonymization approach that combines generalization of QI values with suppression of QS values.

Phase 1 (QI generalization). Given D, and an integer $k >= 1$, we obtain an intermediate dataset D^g that satisfies k-anonymity. There exists many generalization algorithms [2] such as [3] that can be used to obtain a D^g.

Phase 2 (QS suppression). Given D^g, the QI groups that do not satisfy *QS (c,l)-diversity* or *QS t-closeness* will be identified. For these groups, we propose a suppression algorithm to remove proper QS values (items) until the group satisfies *QS (c,l)-diversity* or *QS t-closeness*.

The generalization helps to reduce the work needed in the second step and enhances the overall performance of the system, which we will analyze in Section 3. In fact, when k=1, the system degenerates to the special case in which no generalization is used, and QI groups can be formed by simply putting tuples sharing the same QI values together. For QS suppression, we propose a novel algorithm that uses greedy heuristics but adaptively updates the solution within a time bound.

Cost metric for QS suppression. While various metrics could be used to measure the information loss or cost associated with item removal, in order to maximize the utility of the final output for the suppression phase, we suggest the "inverse weighted cost" metric.

Suppose there is a group of d tuples and each tuple has a row of w_i sensitive terms. When a term is removed from a row with w_i' terms left, we accumulate the total cost by $1/w_i'$, and in the end the cost of all the removed terms is summed up. In this metric, the added cost increases if one keeps removing terms from the same row, which also reflects the impact of the removal to the data utility: the impact is always larger when a sensitive term is removed from a row with less terms left.

Assuming in the end v_i terms are removed from the *ith* tuple, we get the accumulated cost T (normalized by its upperbound):

$$T = \frac{\sum_{i=1}^{d} \sum_{j=w_i - v_i + 1}^{w_i} 1/j}{\sum_{i=1}^{d} \sum_{j=1}^{w_i} 1/j} \quad (1)$$

QS suppression algorithm. A naive approach is to perform a depth-first search as follows. We use a unique number to represent each of the removable QS terms in the to-be-processed QI group. Starting from the root node which indicates the case that no term is removed, we can traverse the entire search tree where each node represents a unique removal pattern until we finish. An example of a search tree ordered lexicologically is shown in Figure 3, where the numbers inside curly braces are the tailsets of the nodes, representing the terms that could be removed in the next depth level.

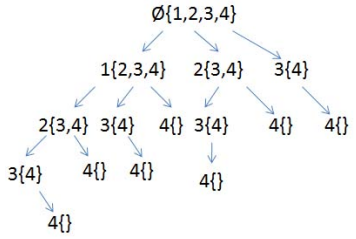


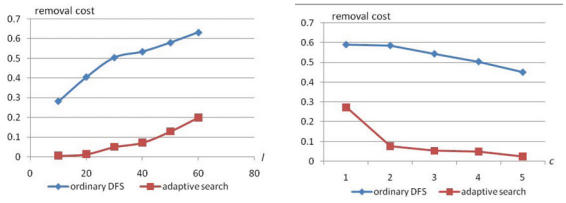
Figure 3: Search Tree for QS Term Suppression

To enhance the performance of the tree traverse, we use adaptive tailset reordering to first explore the most promising branches. Let's first start from the *QS (c,l)-diversity* based algorithm. Imagine that there is a node with a tailset containing the sensitive terms that have not been removed yet, instead of finishing each of its branches one by one, we can tentatively remove a term in the tailset to move one level down, and get: 1) a number l', which is the largest l such that the group represented by the new node satisfies *QS (c,l)-diversity*, 2) the accumulated cost from removing a QS term. By testing all the terms in the tailset in one step, we get the l' and T related with each of these terms. An "effective removal" always features a great increase in l' and a small increase in the cost T, so in a greedy heuristic, it is very natural for us to rank the terms in the tailset by the ratio of the increase in l' and the increase in cost T, i.e., $\Delta l'/\Delta T$. Hence we could reorder the tailset in the decreasing order of $\Delta l'/\Delta T$ for all the tentatively explored nodes, and continue searching the tree recursively in a depth first manner. The first result is obtained when the first node satisfying *QS (c,l)-diversity* is detected.

After the first pattern is found through the above greedy heuristics, the algorithm continues to search for a better solution on other branches of the tree. The result is updated when we find a new node that satisfies *QS (c,l)-diversity* and has a lower cost than the current best result. This process continues until the entire search tree is traversed, so this algorithm generates an optimal solution eventually. This algorithm is greedy in the beginning, but keeps updating the best T so in the end the search is complete. In practice, we can set up a time bound for the algorithm, and mandatorily terminate its execution when time is up. This strategy enables us to find the best solution that we can get in a bounded time period. Similarly, *QS t-closeness* based algorithm can be also optimized with adaptive tailset reordering.

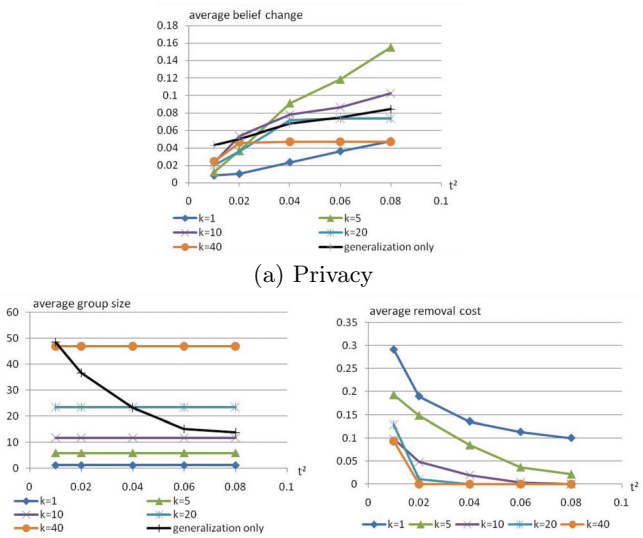
Parameter	Description	Default
k	k-anonymity for generalization	varies
c	QS (c, l)-diversity	1
l	QS (c, l)-diversity	30
t	QS t-closeness	0.4

Figure 4: Experiment Parameters



(a) impact of l (b) impact of c

Figure 5: QS Suppression for *QS (c, l)-diversity*



(a) Privacy

(b) Generalization Cost (c) Suppression Cost

Figure 6: QI Generalization and QS Suppression for *QS t-closeness*

3. EXPERIMENTS

We conducted a set of experiments evaluating the effectiveness and efficiency of the proposed approach. Due to space limitations, we only present the primary results and findings here and refer readers to [6] for more results and more detailed discussions. We implemented the Mondrian generalization algorithm [3] and our suppression algorithm in C++. All experiments are executed on a 2.0 GHz duo core laptop with 2G memory running Windows Vista. The experiments used: 1) a dataset with 3000 tuples augmented from the Adult dataset[1] as D, with 8 QI attributes and 9 synthesized QS terms per tuple, and 2) an external table with 3000 pieces of knowledge labels linked to random QS terms whose appearance in each tuple follows a Poison distribution with $\lambda = 100$.

[1]http://archive.ics.uci.edu/ml/datasets/Adult

QS Suppression algorithm. With the execution time in each experiment bounded by 10 seconds, we first compare our QS suppression algorithm with the baseline DFS algorithm with respect to a set of algorithmic parameters. Figure 4 summarizes the parameters and their default values.

Figure 5 compares the *QS (c, l)-diversity* based QS suppression algorithm against the baseline DFS algorithm with respect to varying l and c. As is shown, the adaptive algorithm significantly outperforms the baseline DFS without tailset reordering on removal cost.

QI generalization and QS suppression. We now present the performance of the overall anonymization approach. For *QS t-closeness* based approach, we tested the impact of different t on privacy and information loss respectively, shown in Figure 6. The X axis is t^2, since we used d_{sq} in the QS suppression algorithm. A direct conclusion is: given a two-phase approach, the average removal cost usually goes down when a larger k is used in generalization; in the extreme case with generalization only approach, we have the smallest term removal cost (zero) and the largest information loss from generalization; and in the extreme case with term removal only approach, we have the largest term removal cost and the least generalization.

4. CONCLUSION AND FUTURE WORKS

We studied the problem of indirect information disclosure through linking quasi-sensitive attributes with external knowledge. Our experiments show that the proposed term removal algorithm is fast and effective, and the two-phase anonymization approach provides bounded privacy guarantees as well as flexible tradeoff between privacy and utility of the data. Our work continues on several directions including alternative information loss metrics such as [8] for term removal, extensions to the algorithm such as additional suppression to cope with minimality attack [7], and potential ways to construct external knowledge base.

Acknowledgements

The research is partially supported by a Career Enhancement Fellowship by Woodrow Wilson Foundation. The authors would like to thank Dr. Mukesh Mohania and his team for kindly sharing the implementation of the ERASE system [1], Dr. Ke Wang for his insightful discussions on anonymizing transactional dataset, and anonymous reviewers for their valuable comments to improve the paper.

5. REFERENCES

[1] V. Chakaravarthy, H. Gupta, P. Roy, and M. Mohania. Efficient techniques for document sanitization. *CIKM*, 2008.
[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4), 2010.
[3] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *IEEE ICDE*, 2006.
[4] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, 2007.
[5] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE*, 2006.
[6] P. Shi and L. Xiong. Protecting quasi-sensitive set-valued data. Technical Report TR-2010-004, Emory University, Department of Mathematics and Computer Science, 2010.
[7] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *VLDB*, 2007.
[8] Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu. Anonymizing transaction databases for publication. In *SIGKDD*, 2008.