2009 IEEE International Conference on e-Business Engineering

Achieving P-Sensitive K-Anonymity via Anatomy

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

Jiuyong Li
School of Computer and Information Science
University of South Australia, Australia
jiuyong.li@unisa.edu.au

David Ross
Faculty of Engineering and Surveying
University of Southern Queensland, Australia
rossd@usq.edu.au

***Abstract*—Privacy-preserving data publishing is to protect sensitive information of individuals in published data while the distortion ratio of the data is minimized. One well-studied approach is the k-anonymity model. Recently, several authors have recognized that k-anonymity cannot prevent attribute disclosure. To address this privacy threat, one solution would be to employ p-sensitive k-anonymity, a novel paradigm in relational data privacy, which prevents sensitive attribute disclosure. p-sensitive k-anonymity partitions the data into groups of records such that each group has at least p distinct sensitive values. Existing approaches for achieving p-sensitive k-anonymity are mostly generalization-based. In this paper, we propose a novel permutation-based approach called *anatomy* to release the quasi-identifier and sensitive values directly in two separate tables. Combined with a grouping mechanism, this approach not only protects privacy, but captures a large amount of correlation in the microdata. We develop a top-down algorithm for computing anatomized tables that obey the p-sensitive k-anonymity privacy requirement, and minimize the error of reconstructing the microdata. Extensive experiments confirm that *anatomy* allows significantly more effective data analysis than the conventional publication methods based on generalization.**

I. Introduction

The problem of privacy-preserving data publishing has received a lot of attention in recent years. Privacy preservation on relational 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 modeling background knowledge of attackers. Most of existing work is formulated in the following context: several organizations, such as hospitals, publish detailed data (also called microdata) about individuals (e.g. medical records) for research or statistical purposes.

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 [10] and privacy breaches caused by AOL search data [3]. Even if identifiers such as names and social security numbers have been removed, the adversary can use linking [9], homogeneity and background attacks [5], [11], [12] to re-identify individual data records or sensitive information of individuals. To overcome the re-identification attacks, the mechanism of k-anonymity was proposed [8], [9]. Specifically, a data set is said to be k-anonymous $(k \geq 1)$ if, on the quasi-identifier (QI) attributes (that is, the maximal set of join attributes to re-identify individual records), each record is identical with at least $(k-1)$ other records. The larger the value of k, the better the privacy is protected.

Although k-anonymity has been well adopted, Traian et al. [11], [12] showed that a k-anonymous table may still have some subtle but severe privacy problems due to the lack of diversity in the sensitive attributes. In particular, they showed that, the degree of privacy protection does not really depend on the size of the QI-groups, which contain tuples that are identical on those attributes. Instead, it is determined by the number of the distinct sensitive values associated with each equivalence class. To overcome the weakness in k-anonymity, they proposed the notion of p-sensitive k-anonymity. Its purpose 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 QI attributes. For instance, Table II is a 2-sensitive 2-anonymous view of Table I since in each QI-group, there are at least two distinct Disease values.

The main approaches used to achieve anonymity in the previous work are based on generalization or suppression [9], [4], [5], which assumes that each attribute has a corresponding conceptual generalization hierarchy or taxonomy tree. A lower level domain in the hierarchy provides more details than a higher level domain. For example, Zip Code 4350 is a lower level domain and Zip Code 435$*$ is a higher level domain. Such hierarchies are suitable for numerical attributes as well. In particular, a hierarchical structure is

978-0-7695-3842-6/09 $26.00 © 2009 Crown Copyright
DOI 10.1109/ICEBE.2009.34

Job	Age	Sex	Zip Code	Disease
clerk	23	M	1100	HIV
manager	37	M	2300	Fever
clerk	61	F	3400	Flu
worker	69	F	4350	HIV
worker	41	M	4435	Flu
technician	58	M	5340	Fever

Table I: Original medical data

Job	Age	Sex	Zip Code	Disease
*	[20-60]	M	[1000-6000]	HIV
*	[20-60]	M	[1000-6000]	Fever
*	[61-70]	F	[1000-6000]	Flu
*	[61-70]	F	[1000-6000]	HIV
*	[20-60]	M	[1000-6000]	Flu
*	[20-60]	M	[1000-6000]	Fever

Table II: 2-sensitive 2-anonymous table

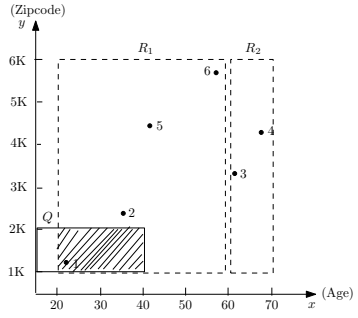

Figure 1: 2D visualization of original and generalized data in Age-Zip code plane.

defined as {value, interval, $*$}, where value is the raw numerical data, interval is the range of the raw data and $*$ is a symbol representing any values. Generalization replaces lower level domain values with higher level domain values. For example, the attribute value of Age 23 in Table I is generalized to the interval [20-60] in Table II.

A. Defects of generalization schema

Although generalization preserves privacy, it often loses considerable information in the microdata, which severely compromises the accuracy of data analysis. Assume that the Table II was released, and that a researcher wants to derive from this table an estimate for the following query:

A: SELECT COUNT(*) FROM medical data
WHERE Disease='fever' AND Age $\leq$ 40 AND Zip Code IN [1000-2000]

To illustrate how to process the query, Figure 1 shows a 2D space, where the x-, y-dimensions are Age and Zip Code, respectively. Each point denotes a tuple in the microdata of Table I. For example, the x-, y-coordinates of point 1 equal the age and Zip code of tuple 1, respectively. Rectangle R_1 (or R_2) is obtained from the generalized values in the first (or second) QI-group in Table II. For instance, the x- (y-) projection of R_1 is the generalized age [20, 60](Zip code [1000, 6000]) of tuples 1,2,5,6. Query A is represented as the shaded rectangle Q, whose projection on the x- (y-) dimension is decided by the range condition Age $\leq$ 40 (1000<Zip code<2000).

Since the researcher sees only R_1 and R_2 (but not the points), s/he answers query A in a way similar to selectivity estimation on a multidimensional histogram [6], as suggested in [4]. Clearly, as R_2 is disjoint with Q, no tuple in the second QI-group can satisfy the query. R_1, however, intersects Q, and hence, is examined as follows. From the Disease in Table II, without additional knowledge, the researcher assumes uniform data distribution in R_1, and computes probability p that a tuple in the QI-group qualifies the range predicates of A as $Area(R_1 \cap R_Q)/Area(R_1)$=0.1. This value leads to an approximate answer 0.2, which, however, is five times smaller than actual query result 1.

The gross error is caused by the fact that the data distribution in R_1 significantly deviates from uniformity. Nevertheless, given only the generalized table, we cannot justify any other distribution assumption. This is an inherent problem of generalization, which prevents an analyst from correctly understanding the data distribution inside each QI-group.

B. The anatomy approach

The idea of the anatomy approach is that if two tables with a join attribute are published, the join of the two tables can be lossy and this lossy join helps to conceal the private information [15].

We could like to make use of this idea to derive a new mechanism for achieving a privacy principle p-sensitive k-anonymity. For example Table IV is a 2-sensitive 2-anonymous view of Table III. From this table, we can generate a temporary table shown in Table V. For each QI-group Q in the anonymized table, we assign a unique identifier (ID) to Q and also to all tuples in Q. Then, we attach the correspondence ID to each tuple in the original raw table and form a new table temporary. From the temporary table, we can generate two separate tables, Table VI and VIII. The two tables share the attribute of Group ID. If we join these two tables by the Group ID, it is easy to see that the join is lossy and it is not possible to derive the Table V after the join. The result of joining is shown in Table X. From the anatomy, each sensitive value is linked to at least 2 values. Therefore, the required individual privacy 2-diversity is guaranteed (formal proof given in Section 4). Also, in the joined table, for each individual, there are at least 2 individuals that are linked to the same bag B of sensitive values, such that in terms of the sensitive values, they are not distinguishable. For example, the first record in the original table (QI = (clerk, 1975, 4350)) is linked to bag

Job	Birth	Zip Code	Disease
clerk	1975	4350	HIV
manager	1955	4350	Flu
clerk	1955	5432	Flu
worker	1955	5432	Fever
worker	1975	4350	Flu
technician	1940	4350	Fever

Table III: Raw medical data

Job	Birth	Zip Code	Disease
White-collar	*	4350	HIV
White-collar	*	4350	Flu
*	1955	5432	Flu
*	1955	5432	Fever
Blue-collar	*	4350	Flu
Blue-collar	*	4350	Fever

Table IV: 2-sensitive 2-anonymous table

Job	Birth	Zip Code	Disease	ID
clerk	1975	4350	HIV	1
manager	1955	4350	Flu	1
clerk	1955	5432	Flu	2
worker	1955	5432	Fever	2
worker	1975	4350	Flu	3
technician	1940	4350	Fever	3

Table V: Temporary table

$B = \{HIV, flu\}$. We find that the second individual (QI = (manager, 1955, 4350)) is also linked to the same bag B of sensitive values. This is the goal of p-sensitive k-anonymity for the protection of sensitive values.

C. Contribution and paper organization

[11] proposed to generate one generalized table which satisfies p-sensitive k-anonymity by full domain generalization. Due to the defects of generalization-based approach, we propose a novel permutation-based approach, called "anatomy", with which the privacy protection for p-sensitive k-anonymity can be achieved by releasing the quasi-identifier and sensitive values directly in two separate tables. In the two tables, one table contains the undisturbed non-sensitive values and the other table contains the undisturbed sensitive values. We show that the results are better than previous approaches in the experiments.

The rest paper is organized as follows. In Section 2, we introduce some basic concepts for privacy protection. In Section 3, we discuss p-sensitive k-anonymity model and illustrate how to anonymize the microdata with p-sensitive k-anonymity requirement by the permutation-based approach. We propose an efficient top-down algorithm for p-sensitive k-anonymity with anatomy in Section 4. The extensive experimental studies are included in Section 5. We conclude the paper in Section 6.

II. Preliminaries

Let T be the initial microdata table and T' be the released microdata table. T' consists of a set of tuples over an attribute set. The attributes characterizing microdata are classified into the following three categories.

• *Identifier attributes* that can be used to identify a record such as Name and Medicare card.

• *Quasi-identifier (QI) attributes* that may be known by an intruder, such as Zip code and Age. QI attributes are presented in the released microdata table T' as well as in the initial microdata table T.

• *Sensitive attributes* that are assumed to be unknown to an intruder and need to be protected, such as Health Condition or ICD9Code [1]. Sensitive attributes are presented both in T and T'.

[1]available at http://www.icd9code.com/

In what follows we assume that the identifier attributes have been removed and the quasi-identifier and sensitive attributes are usually kept in the released and initial microdata table. Another assumption is that the value for the sensitive attributes are not available from any external source. This assumption guarantees that an intruder can not use the sensitive attributes to increase the chances of disclosure. Unfortunately, an intruder may use record linkage techniques [13] between quasi-identifier attributes and external available information to glean the identity of individuals from the modified microdata. To avoid this possibility of privacy disclosure, one frequently used solution is to modify the initial microdata, more specifically the quasi-identifier attributes values, in order to enforce the k-anonymity property.

Definition 1 (Quasi-Identifier): A quasi-identifier (QI) is a minimal set Q of attributes in microdata table T that can be joined with external information to re-identify individual records (with sufficiently high probability).

Definition 2 (k-anonymity): The modified microdata table T' is said to satisfy k-anonymity if and only if each combination of quasi-identifier attributes in T' occurs at least k times.

A QI-group in the modified microdata T' is the set of all records in the table containing identical values for the QI attributes. There is no consensus in the literature over the term used to denote a QI-group. This term was not defined when k-anonymity was introduced [7], [10]. More recent papers use different terminologies such as equivalence class [14] and QI-cluster [11], [12].

For example, let the set {Job, Birth, Zip Code} be the quasi-identifier of Table III. Table IV is one 2-anonymous view of Table III since there are two QI-groups and the size of each QI-group is at least 2. So k-anonymity can ensure that even though an intruder knows a particular individual is in the k-anonymous microdata table T, s/he can not infer which record in T corresponds to the individual with a probability greater than $1/k$.

III. p-sensitive k-anonymity with anatomy

Let us re-examine the objectives of p-sensitive k-anonymity. With k-anonymity, we want to make sure that when an individual is mapped to some sensitive values, at least $k-1$ other individuals are also mapped to the same sensitive values. With p-sensitive, we want to make sure that the diversity of the sensitive values is at least p. For example, consider an individual with (clerk, 1975, 4350) in

Job	Birth	Zip Code	Group ID
clerk	1975	4350	1
manager	1955	4350	1
clerk	1955	5432	2
worker	1955	5432	2
worker	1975	4350	3
technician	1940	4350	3

Table VI: Non-sensitive table (NSS)

Original QI values	Group ID

Table VII: NSS table

Disease	Group ID
HIV	1
Flu	1
Flu	2
Fever	2
Flu	3
Fever	3

Table VIII: Sensitive table (SS)

Group ID	Sensitive attributes

Table IX: SS table

Job	Birth	Zip Code	Disease	Group ID
clerk	1975	4350	HIV	1
manager	1955	4350	HIV	1
clerk	1975	4350	Flu	1
manager	1955	4350	Flu	1
clerk	1955	5432	Flu	2
worker	1955	5432	Flu	2
clerk	1955	5432	Fever	2
worker	1955	5432	Fever	2
worker	1975	4350	Flu	3
technician	1940	4350	Flu	3
worker	1975	4350	Fever	3
technician	1940	4350	Fever	3

Table X: Resulting join table

Table III, with 2-sensitive 2-anonymity property, since s/he is mapped to the first and the second tuple in Table IV, and the diseases are mapped into the same bag B={HIV, Flu}. There is another individual with (manager, 1955, 4350) in Table III that is mapped to the same bag B as well. The p-sensitive k-anonymity property ensures that at least p distinct values exist in each sensitive values set B. For instance, with $p = 2, k = 2$, B contains HIV and Flu. Based on this, we give out the definition of p-sensitive k-anonymity as follows:

Definition 3 (p-sensitive k-anonymity): The modified microdata table T' satisfies p-sensitive k-anonymity property if it satisfies k-anonymity, and for each QI-group in T', the number of distinct values for each sensitive attribute occurs at least p times within the same QI-group.

Suppose we form an anonymous table in which some Quasi-identifier(QI) values are generalized. As we discussed before, in the anonymous table, each set of tuples with the same QI values forms a QI-group. However, instead of publishing one single table T with the generalized values, there is the possibility of separating the sensitive attribute from the non-sensitive attributes and generate two tables by projecting these two sets of attributes. Tuples in the two tables are linked if they belong to the same QI-group in T. Hence we can publish two tables: (1) one table, called non-sensitive table (NSS table), containing all the non-sensitive attributes together with QI-group IDs in T, and (2) the other table, called sensitive table (SS table), containing the QI-group ID and the sensitive attributes. The released tables are annotated with the remark that each tuple in each of the two published table corresponds to one record in the original single table. This is to ensure that a user will not mistakenly join the two tables and assume that the join result corresponds to the original table.

The schema of the non-sensitive table (NSS table) and the sensitive table (SS table) are shown in Table VII and IX, where Group ID corresponds to QI-group ID. An example is shown in Table VI and VIII. The following theorem proves that the resulting table indeed satisfies the p-sensitive k-anonymity property.

Theorem 1: The resulting published tables NSS and SS satisfy p-sensitive k-anonymity property.

Proof: Given the QI information of individuals in a table T_I (which we assume that an attacker may possess) and the anonymous Table T_A (e.g. Table IV). We can "join" the two tables by matching each QI in T_A to its anonymous QI-group and obtain a table T_{IA}. Since T_A satisfies p-sensitive k-anonymity, when the QI values of an individual is linked to a set S of values in the sensitive attribute, at least $k-1$ other QI's values of other individuals are also linked to S. In addition, distinct sensitive values in S are at least p.

Now, suppose the adversary is given tables NSS and SS. Equipped with only table T_I, an adversary must join the tables NSS and SS on their common attribute in an attempt to link the QI's values to the sensitive values. Let the join result be table T'_A, such as Table X. Consider any QI-group with group ID X. Let S_X be the set of sensitive values that X is linked to in T_A and suppose there are a tuples in T_A belonging to X. In Table T'_A, there will be p^2 tuples generated for X and S_X becomes S'_X and each entry in S_X is duplicated p times in S'_X. In the p^2 tuples in T'_A, each original QI's values in the given table T will now be linked to the set S'_X. Besides, p individuals are involved in X, and each is linked to S_X. The number of different sensitive value in S'_X is the same as that in S_X in T_{IA}. Hence, the tables NSS and SS release no more information as the table T_A in terms of the linkage of an individual to a set S of sensitive values and in terms of the number of distinct sensitive values in S.

This shows that the privacy protection provided by the single anonymous table T_A is no stronger than that provided by the NSS and SS tables in terms of p-sensitive k-anonymity. Since T_A satisfies p-sensitive k-anonymity, tables NSS and SS also satisfy p-sensitive k-anonymity. ■

For better understanding, the example shown in Tables IV to X demonstrates the ideas of the proof above. If we publish Tables VI and VIII, we can achieve similar privacy preservation objectives as we publish Table IV only.

IV. Experiments

This section empirically evaluates the effectiveness and efficiency of the anatomy approach for p-sensitive k-anonymity. We first clarify the settings of our experi-

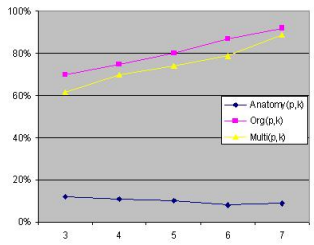

Figure 2: Query accuracy vs. the number of QI d

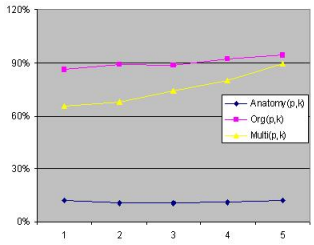

Figure 3: Query accuracy vs. query dimensionality qd

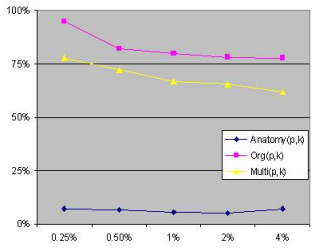

Figure 4: Query accuracy vs. expected sensitivity s

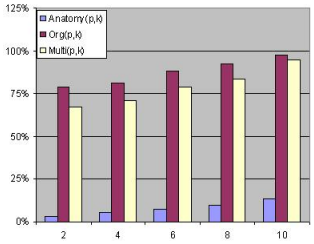

Figure 5: Query accuracy vs. parameter p

ments. Then, we compare anatomy with two other existing anonymization methods in terms of their effectiveness for data analysis and computation cost.

Experiment Setup: We adopted the adult data set for the experiment, which can be downloaded in the UCIrvine Machine Learning Repository[2]. We eliminated the records with unknown values in this data set. The resulting data set contains 45,222 tuples. Nine of the attributes were chosen in our experiments, as shown in Table XI. From adult dataset, we create a set of microtables, in order to examine the influence of dimensionality. The set contains 5 tables, denoted as adult-3, $\cdots$, adult-7, respectively. Specifically, adult-d $(3 \leq d \leq 7)$ treats the first d attributes in Table XI as the QI-attributes, and the attribute Occupation as the sensitive attribute A^s. For example, adult-4 is 5D, and contains QI-attributes Age, Work Class, Education, Martial Status and the sensitive attribute Occupation. Furthermore, to study the impact of cardinality, we generate multiple versions of each adult-d with various cardinalities n, by randomly sampling n tuples from the whole adult-d with 45,222 tuples.

Attribute	Distinct Values	Generalization	Height
Age	74	5-,1-,20-year ranges	4
Work Class	7	Taxonomy Tree	3
Education	16	Taxonomy Tree	4
Martial Status	7	Taxonomy Tree	3
Race	5	Taxonomy Tree	2
Sex	2	Suppression	1
Native Country	41	Taxonomy Tree	3
Salary Class	2	Suppression	1
Occupation	14	Taxonomy Tree	2

Table XI: Features of Adult Data

We implemented our proposed algorithm, the p-sensitive k-anonymity based privacy preservation by anatomy. Let us denote it as $Anatomy(p, k)$. We compared the proposed top-down algorithm with the original algorithm of p-sensitive k-anonymity [11] which generalizes the QI and forms one generalized table only. Let us denote the algorithm by $Org(p, k)$. We also compare our algorithm with the state-of-the-art algorithm in [4], which adopts multi-dimension recoding. We denote it as $Multi(p, k)$. Recall that each QI value is generalized and the last two columns of Table XI describes how these QI attributes are generalized.

By default, we set $p = 2, k = 20$. For adult-d tables, p distributes from 2 to 10, and k ranges from 5 to 30. The utility of an anonymization technique is evaluated in terms of effectiveness for aggregate queries. The effectiveness of aggregate query is defined to be its average relative error in answering a query of the following form.

SELECT COUNT(*)
FROM Unknown-Microdata
WHERE $pred(A_1^{qi})$ AND $\cdots$ AND $pred(A_{qd}^{qi})$ AND $pred(A^s)$

In the above query, Unknown-Microdata is an original data set or an anonymized data set. The *query dimensionality* qd denotes the number of QI attributes to be queried and A^s denotes the sensitive attribute. For instance, if the microdata is adult-3 and $qd = 2$, then $\{A_1^{q_i}, A_2^{q_i}\}$ is a random 2-sized subset of {Age, Work Class, Education}. For any attribute A, the predicate $pred(A)$ has the form ($A = x_1$ OR $A = X_2$ OR $A = x_b$), where x_i is a random value in the domain of A, for $1 \leq i \leq b$. The value of b depends on the *expected query selectivity* s: $b = \lceil |A| \cdot s^{1/(qd+1)} \rceil$, where $|A|$ is the domain size of A. If the value of s is set higher, the selection conditions in $pred(A)$ will be more.

We compare the anonymized tables generated by different algorithms in terms of average relative error, which is defined as follows. We perform the aggregate query with the original data set, called Original. That is,

SELECT COUNT(*)
FROM Original
WHERE $pred(A_1^{qi})$ AND $\cdots$ AND $pred(A_{qd}^{qi})$ AND $pred(A^s)$

Let us call the count obtained above act. We execute the aggregate query with the anonymized data set as follows. As algorithm $Anatomy(p, k)$ generates two tables, namely NSS and SS, we perform the query as follows.

SELECT COUNT(*)
FROM SS
WHERE SS.Group ID in (SELECT NSS.Group ID FROM NSS, WHERE $pred(A_1^{qi})$ AND $\cdots$ AND $pred(A_{qd}^{qi})$ AND $pred(A^s)$)

[2]http://www.ics.uci.edu/mlearn/MLRepository.html.

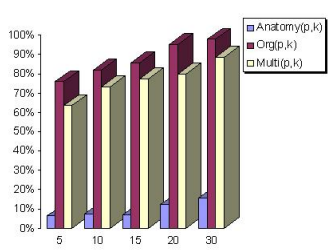

Figure 6: Query accuracy vs. parameter k

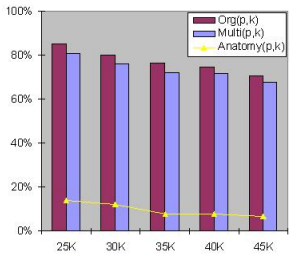

Figure 7: Query accuracy vs. dataset cardinality n

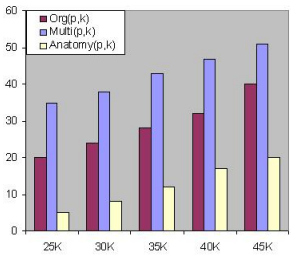

Figure 8: Running time vs. dataset cardinality n

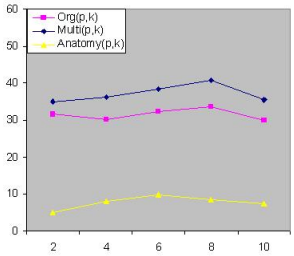

Figure 9: Running time vs. parameter p

Let us call the count obtained above est. As algorithms $Org(p,k)$ and $Multi(p,k)$ generate one anonymized table, we perform the first query by replacing Unknown-Microdata with the anonymized or generalized data. Then, we define the relative error to be $|act-est|/act$, where act is its actual count derived from the original, and est the estimated count computed from the anonymized table.

Table XII summarizes the parameters of our experimentation, as well as their values examined. The values in bold are the defaults.

Parameter	Value
p	**2**,4,6,8,10
k	5,10,15,**20**,30
cardinality n	**25K**,30K,35K,40K,45K
number of QI-attributes d	3,4,**5**,6,7
query dimensionality qd	1,2,$\cdots$,d
expected sensitivity s	0.25%,**0.5%**,1%,2%,4%

Table XII: Parameter and tested values

Effectiveness of data analysis: This section compares alternative solutions on their accuracy in count analysis. The first set of experiments examines the effects of d. Figure 2 plots the error of all methods as a function of d, for datasets adult-d. From the Figure 2, we can see that $Anatomy(p,k)$ gives a lower average relative error compared with the other two algorithms $Org(p,k)$, $Multi(p,k)$. This is because algorithm $Anatomy(p,k)$ does not generalize the table but algorithm $Org(p,k)$ generalize the table, which makes the average relative error higher, and its accuracy decays severely as the dimensionality increases, which confirm the analysis in [1].

Next, we examine alternative techniques on queries involving different number qd of QI-attributes. Figure 3 demonstrates the results for adult-5. Again, the error of anatomy is consistently small, whereas generalization-based approaches are again the worst techniques. Figure 4 present the error as a function of query selectivity s. It can be found that the average relative error of all three algorithms decreases when s increases. This is because, when s is larger, each attribute in the aggregate query involves more value matches. That means the actual count is larger. Note that the actual count is the denominator of the average relative error. Besides, if the generalized values in the anonymized table match more aggregate values in the query, the estimated count will be more accurate. Thus, the overall average relative error decreases when s increases.

Figure 5 examines the accuracy as p varies. As expected, the error of anatomy-based and generalization-based increases with p, because more data distortion is needed in order to enforce stronger privacy control. Similar results are plotted by varying k in Figure 6. Finally, Figure 7 presents the results under different dataset cardinalities n. We found that the average relative error of all three algorithms decreases slightly when n increases. This is because, when n is larger, there is more chance that a tuple can be matched with an existing tuple in the data without much generalization. Similarly, algorithm $Anatomy(p,k)$ gives a lower average relative error compared with algorithm $Org(p,k)$ and $Multi(p,k)$. The effectiveness of each method remains fairly stable at all cardinalities.

Efficiency of the algorithms: Having tested the effectiveness of anatomy-based approach for data analysis, we proceed to evaluate its efficiency. Figure 8 compares the time of anonymization required by $Anatomy(p,k)$, $Org(p,k)$ and $Multi(p,k)$ on the dataset with different cardinalities n. The computing overhead of $Anatomy(p,k)$ is less than the other two generalization-based approaches, since there is no generalization cost. Similar results are shown in Figure 9, where the the experiments are carried out shows by varying k for all three algorithms.

V. Conclusion and Future Work

In this paper, we proposed a p-sensitive k-anonymity based privacy preservation mechanism that reduce information loss by the anatomy approach. Instead of publishing one generalized table, we generate two tables with a sharing attribute called Group ID, which corresponds to a unique identifier of a "QI-group". One table contains the detailed information of the quasi-identifier and Group ID, and the other table contains Group ID and the sensitive attribute. By avoiding the generalization of the quasi-identifier in the first table, we achieve less information loss. We conducted

extensive experiments and verified the improvement on effectiveness on data analysis and efficiency in execution time.

This work also initiates several directions for future investigation. For example, in this article, we focused on the case where there is a single sensitive attribute; extending our technique to multiple sensitive attributes is an interesting topic. Another direction is to apply this permutation-based approach to other privacy principles, and make comprehensive experimental studies to show its effectiveness and efficiency.

ACKNOWLEDGEMENT

We would like to thank anonymous reviewers useful comments. This research is supported by Australian Research Council (ARC) grant DP0774450 and DP0663414.

REFERENCES

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In Proc. of Very Large Data Bases (VLDB), pages 901C909, 2005.

[2] 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.

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

[4] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In ICDE, 2006.

[5] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. p-sensitive k-anonymity: Privacy beyond k-anonymity. In ICDE, 2006.

[6] N. Thaper, S. Guha, P. Indyk, and N. Koudas. Dynamic multidimensional histograms. In SIGMOD, pages 428C439, 2002.

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

[8] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing Information. PODS 1998.

[9] 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.

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

[11] 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.

[12] T. M. Truta, A. Campan and P. Meyer. Generating Microdata with p-sensitive k-anonymity Property. *SDM 2007*: 124-141

[13] W. E. Winkler. Advanced Methods for Record Linkage, *Proceedings of the Section on Survey Research Methods, American Statistical Society*, 467-472

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

[15] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In VLDB, 2006