

(p^+, α) -sensitive k -anonymity: A new enhanced privacy protection model

Xiaoxun Sun Hua Wang
Department of Mathematics & Computing
University of Southern Queensland
Queensland, Australia
{sunx, wang}@usq.edu.au

Traian Marius Truta
Department of Computer Science
Northern Kentucky University
Highland Heights, KY, U.S.A
trutat1@nku.edu

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

Ping Li
School of Economics and Management
Beihang University
Beijing 10083, China
lipingxx@126.com

Abstract

Publishing data for analysis from a microdata table containing sensitive attributes, while maintaining individual privacy, is a problem of increasing significance today. The k -anonymity model was proposed for privacy preserving data publication. While focusing on identity disclosure, k -anonymity model fails to protect attribute disclosure to some extent. Many efforts are made to enhance the k -anonymity model recently. In this paper, we propose a new privacy protection model called (p^+, α) -sensitive k -anonymity, where sensitive attributes are first partitioned into categories by their sensitivity, and then the categories that sensitive attributes belong to are published. Different from previous enhanced k -anonymity models, this model allows us to release a lot more information without compromising privacy. We also provide testing and heuristic generating algorithms. Experimental results show that our introduced model could significantly reduce the privacy breach.

1. Introduction

With the rapid growth in database, networking, and computing technologies, a large amount of personal data can be integrated and analyzed digitally, leading to an increased use of data-mining tools to infer trends and patterns. This has raised universal concerns about protecting the privacy of individuals.

Many data holders publish their microdata for different purposes. However, they have difficulties in releasing in-

formation which does not compromise privacy. The traditional approach of releasing the data tables without breaching the privacy of individuals in the table is to de-identify records by removing the identifying fields such as name, address, and social security number. However, joining this de-identified table with a publicly available database (like the voters database) on attributes like race, age, and zip code (usually called quasi-identifier) can be used to identify individuals.

In order to protect privacy, Sweeney [14] proposed the k -anonymity model, where some of the quasi-identifier fields are suppressed or generalized so that, for each record in the modified table, there are at least $k - 1$ other records in the modified table that are identical to it along the quasi-identifier attributes. For the Table 1, Table 5 shows a 4-anonymous view corresponding to it. The sensitive attributes (Health Condition) is retained without change in this example.

In recent years, numerous algorithms have been proposed for implementing k -anonymity via generalization and suppression. Samarati [10] presents an algorithm that exploits a binary search on the domain generalization hierarchy to find minimal k -anonymous table. Sun *et al.* [11] recently improve his algorithm by integrating the hash-based technique. Bayardo and Agrawal [2] presents an optimal algorithm that starts from a fully generalized table and specializes the dataset in a minimal k -anonymous table, exploiting ad hoc pruning techniques. LeFevre *et al.* [5] describes an algorithm that uses a bottom-up technique and a priori computation. Fung *et al.* [3] present a top-down heuristic to make a table to be released k -anonymous. As to the theoretical results, Meyerson and Williams [8] and Aggarwal *et al.* [1] proved the optimal k -anonymity is NP-

ID	Age	Country	Zip Code	Health Condition
1	27	USA	14248	HIV
2	28	Canada	14207	HIV
3	26	USA	14246	Cancer
4	25	Canada	14249	Cancer
5	41	China	13053	Hepatitis
6	48	Japan	13074	Phthisis
7	45	India	13064	Asthma
8	42	India	13062	Heart Disease
9	33	USA	14242	Flu
10	37	Canada	14204	Flu
11	36	Canada	14205	Flu
12	35	USA	14248	Indigestion

Table 1: Microdata

ID	Age	Country	Zip Code	Health Condition
1	<30	America	142**	HIV
2	<30	America	142**	HIV
3	<30	America	1424*	Cancer
4	<30	America	1424*	Cancer
5	>40	Asia	130**	Hepatitis
6	>40	Asia	130**	Phthisis
7	>40	Asia	130**	Asthma
8	>40	Asia	130**	Heart Disease
9	3*	America	1424*	Flu
10	3*	America	142**	Flu
11	3*	America	142**	Flu
12	3*	America	1424*	Indigestion

Table 2: One 2-anonymous view of Table 1

hard (based on the number of cells and number of attributes that are generalized and suppressed) and describe approximation algorithms for optimal k -anonymity. Sun *et al.* [12] proved that k -anonymity problem is also NP-hard even in the restricted cases, which could imply the results in [1, 8] as well.

While focusing on identity disclosure, k -anonymity model fails to protect attribute disclosure [4]. Several models such as p -sensitive k -anonymity [15], l -diversity [7], (α, k) -anonymity [19] and t -closeness [6] were proposed in the literature in order to deal with the problem of k -anonymity. The work presented in this paper is highly inspired by [15]. The main contribution of [15] is to introduce the p -sensitive k -anonymity property, which requires, in addition to k -anonymity, that for each group of tuples with identical combination of quasi-identifier values, the number of distinct sensitive attributes values must be at least p . However, depending on the nature of the sensitive attributes, even p -sensitive property still permits the information to be disclosed. We identify in this paper, situations when p -sensitive property is not enough for privacy protection and we propose a solution to overcome this identified problem: (p^+, α) -sensitive k -anonymity model and a heuristic algorithm to enforce this property.

The paper is organized as follows. In Section 2, we introduce some basic concepts in k -anonymity. Our new privacy protection model is defined in Section 3. The testing and heuristic algorithms for (p^+, α) -sensitive k -anonymity property is presented in Section 4. Our experimental study is included in Section 5. Finally, we conclude the paper in Section 6.

2. Concepts and problem definition

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. Sensitive attributes are presented both in T and T' .

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 [18] 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

Name	Age	Country	Zip Code
Rick	26	USA	14246
Hassen	45	India	13064
Rudy	25	Canada	14249
Yamazaki	48	Japan	13074

Table 3: External available information

Category ID	Sensitive attribute values	Sensitivity
One	HIV, Cancer	Top Secret
Two	Phthisis, Hepatitis	Secret
Three	Heart Disease, Asthma	Less Secret
Four	Flu, Indigestion	Non Secret

Table 4: Categories of Health Condition

ID	Age	Country	Zip Code	Health Condition
1	<30	America	142**	HIV
2	<30	America	142**	HIV
3	<30	America	142**	Cancer
4	<30	America	142**	Cancer
5	>40	Asia	130**	Hepatitis
6	>40	Asia	130**	Phthisis
7	>40	Asia	130**	Asthma
8	>40	Asia	130**	Heart Disease
9	3*	America	142**	Flu
10	3*	America	142**	Flu
11	3*	America	142**	Flu
12	3*	America	142**	Indigestion

Table 5: 2-sensitive 4-anonymous Microdata

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 [10, 14]. More recent papers use different terminologies such as equivalence class [19] and QI-cluster [16].

For example, let the set {Age, Country, Zip Code} be the quasi-identifier of Table 1. Table 2 is one 2-anonymous view of Table 1 since there are five 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 $1/k$.

The k -anonymity property ensures protection against identity disclosure, i.e. the identification of an entity (person, institution). However, as we will show next, it does not protect the data against attribute disclosure, which occurs when the intruder finds something new about a target entity.

Still consider the modified 2-anonymous table (Table 2), where the set of quasi-identifier is composed of {Age, Country, Zip Code} and Health Condition is the sensitive attribute. As we discussed above, identity disclosure does not happen in this modified microdata. However, assuming that external information in Table 3 is available, attribute disclosure can take place. If the intruder knows that in the modified table (Table 2) the Age attribute was modified to '<30', he can deduce that both Rick and Rudy have Cancer, even he does not know which record, 3 or 4, corresponding to which person. This example shows that even if k -anonymity can well protect identity disclosure, sometimes it fails to protect against sensitive attribute disclosure.

To deal with this problem in privacy breach, the p -sensitive k -anonymity model was introduced in [15]. A similar privacy model, called l -diversity, is described in [7].

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.

Sometimes, the domain of the sensitive attributes, especially the categorical ones, can be partitioned into categories according to the sensitivity of attributes. For example, in medical datasets Table 1, the Health Condition attribute can be classified into four categories (see Table 4). The different types of diseases are organized in a category domain. The attribute values are very specific, for example they can represent HIV or Cancer, which are both Top Secret information of the individuals. In the case that the initial microdata contains specific sensitive attributes like Health Condition, the data owner can be interested in protecting not only these most specific values, but also the category that the sensitive values belong to. For example, the information of a person who affected with Top Secret needs to be protected, no matter whether it is HIV or Cancer. If we modify the microdata to satisfy p -sensitive k -anonymity property, it is possible that in a QI-group with p distinct sensitive attribute values, all of them belong to the same pre-defined category. For instance, the values {HIV, HIV, Cancer, Cancer} in one QI-group in Table 5 all belong to Top Secret category. To avoid such situations, we introduce our new (p^+, α) -sensitive k -anonymity model, which is capable of protecting sensitive values as well as the categories that are considered sensitive.

3. (p^+, α) -sensitive k -anonymity model

Let S be a categorical sensitive attribute we want to protect against attribute disclosure. First, we sort out the values of S according to their sensitivity, forming an ordered value domain D , and then partition the attribute domain into m -categories (S_1, S_2, \dots, S_m) , such that $S = \cup_{i=1}^m S_i$, $S_i \cap S_j = \emptyset$ (for $i \neq j$). $S_i \leq S_j$ means S_i is more sensitive than the S_j (for $1 \leq i \leq j \leq m$). For example, Consider the Health Condition $S = \{HIV, Cancer, Phthisis, Hepatitis, Heart Disease, Asthma, Flu, Indigestion\}$ in Table 1, it has

ID	Age	Country	Zip Code	Health Condition	Category ID
1	<40	America	1424*	HIV	One
4	<40	America	1424*	Cancer	One
9	<40	America	1424*	Flu	Four
12	<40	America	1424*	Indigestion	Four
5	>40	Asia	130**	Hepatitis	Two
6	>40	Asia	130**	Phthisis	Two
7	>40	Asia	130**	Asthma	Three
8	>40	Asia	130**	Heart Disease	Three
2	<40	America	1420*	HIV	One
3	<40	America	1420*	Cancer	One
10	<40	America	1420*	Flu	Four
11	<40	America	1420*	Flu	Four

Table 6: $(2^+, 2)$ -sensitive 4-anonymous Microdata

been partitioned into four categories according to the sensitivity of the diseases (Table 4), where S_1 (Top Secret) is the most sensitive and S_4 (Non Secret) is the least one.

Furthermore, in order to measure the distance between two categories (attributes) and the degree that sensitive attribute values contribute to one QI-group, we introduce the following ordinal metric system.

Let $D(S)$ denote a categorical domain of an attribute S and $|D(S)|$ be the total number of categories in domain $D(S)$. The normalized distance between two categories S_i and S_j of the attribute S with $S_i \leq S_j$ is:

$$d(S_i, S_j) = \frac{|S_i|S_i \leq S_i < S_j|}{|D(S)| - 1}$$

The distance between two sensitive attribute values is equal to the distance between the categories that they fall into.

Moreover, we put an ordinal weight to each category to represent the degree that each specific sensitive value from S protects against the disclosure of values from sensitive categories.

Let $D(S) = \{S_1, S_2, \dots, S_k\}$ denote a partition of categorical domain of an attribute S and let $weight(S_i)$ denote the weight of category S_i . Then,

$$\begin{cases} weight(S_1) = 0, \\ weight(S_i) = \frac{i-1}{k-1}; \quad 1 < i < k \\ weight(S_k) = 1, \end{cases} \quad (1)$$

Note that the weight of the specific sensitive value is equal to the weight of the category that the specific value belongs to. The weight of the QI-group is the total weight of each specific sensitive values that the QI-group contains.

We illustrate these concepts by taking Table 6 as an example. Given the partition of sensitive attributes as shown in Table 4 and four corresponding values set $A = \{\text{Cancer, Phthisis, Asthma, Flu}\}$. The distance between Cancer (S_1) and

Algorithm 1: $check(T, p, k)$

Input: a microdata table T , a user defined classification of sensitive attributes domain $D = (S_1, S_2, \dots, S_m)$ and integer α, p, k ($2 \leq p \leq k$);

Output: True (T is (p^+, α) -sensitive k -anonymous); False (Otherwise)

1. If T is k -anonymous
2. Condition = True;
3. Compute the $weight(S_i)$ for $i = 1, 2, \dots, m$ according to (1).
4. For each QI-group in T
5. Let d be the number of different weight values.
5. Let α' be total weight of the QI-group.
6. If $d < p$ or $\alpha' < \alpha$
7. Condition is False;
8. Else
9. Break loop;
10. Condition is True.
11. Condition is False;

Flu (S_4) is $3/3=1$, while the distance between Phthisis (S_2) and Asthma (S_3) is $1/3$. According to (1), $weight(S_1) = 0$, $weight(S_2) = 1/3$ and $weight(Asthma) = 2/3$, $weight(Flu) = 1$, the total weight of A is $0+1/3+2/3+1=2$.

Definition 4. $((p^+, \alpha)$ -sensitive k -anonymity): The modified microdata table T' satisfies (p^+, α) -sensitive k -anonymity property if it satisfies k -anonymity, and each QI-group has at least p distinct categories of the sensitive attribute and its total weight is at least α .

Table 6 is a $(2^+, 2)$ -sensitive 4-anonymous view of Table 1. As you can see, for example, the records 1,4,9 and 12 belong to one QI-group in which the Health Condition is not that easy to be referred since they belong to two different categories with its total weight 2. Compared with previous 2-sensitive 4-anonymity model (See Table 5), our new model could overcome the shortcomings of previous models and significantly reduce the possibility of leaking privacy. Different from previous models regardless of p -sensitive k -anonymity or l -diversity, here, instead of publishing original specific sensitive attributes, we publish the categories that the sensitive values belong to.

4. The algorithm

In this section, we first present an algorithm to test the (p^+, α) -sensitive k -anonymity property for microdata table, and then we present a simple heuristic switching algorithm to generate a (p^+, α) -sensitive k -anonymous microdata table.

The Algorithm 1 is used in the process of checking if the given microdata table satisfying (p^+, α) -sensitive k -anonymity property. Algorithm 2 is to generate a (p^+, α) -sensitive k -anonymous table. First, we generate a p -sensitive k -anonymous table using algorithm described in [15] (Line 1). Then we check if the generated p -sensitive

Algorithm 2: Switching algorithm

Input: a microdata table T , a user defined classification of sensitive attributes domain $D = \langle S_1, S_2, \dots, S_m \rangle$ and integer α, p, k ($2 \leq p \leq k$);

Output: a (p^+, α) -sensitive k -anonymous microdata table T^* .

1. Generate a minimal p -sensitive k -anonymous table T' .
2. If $\text{check}(T', p, k) = 1$
3. Return T' .
4. Else
5. While some QI -group in T' contain less than p different weight values
5. or the total weight is less than α
6. Switch tuples to form new QI -groups and a temporary table T^* .
7. Anonymize each new QI -group in T^* .
8. Return T^* .

k -anonymous table already satisfies (p^+, α) -sensitive k -anonymity property (Line 2). If not, we target on the QI -groups with $d < p$ or $\alpha' < \alpha$ (Line 5) and switch tuples among QI -groups until d is at least p and α' is at least α in each new formed QI -group (Line 6). Finally, we make each QI -group k -anonymous by generalization (Line 7).

To explain our algorithms, consider Table 5 which satisfies 2-sensitive 4-anonymity property, not $(2^+, 2)$ -sensitive 4-anonymity property, since the Health Conditions in the first QI -group (color Blue) and the last QI -group (Color Red) are in the same category respectively. Intuitively, we can simply switch the tuples between these two QI -groups to obtain the microdata table that satisfies $(2^+, 2)$ -sensitive 4-anonymity property. As Table 6 is obtained by switching tuples 2 and 3 in the first QI -group with tuples 9 and 12 in the third QI -group, and then make it satisfy 4-anonymity property by generalization.

5. Experimental results

In our experiments, we used the *adult* database from the UCI Machine Learning Repository [9]. This database has been used by many researchers and become the benchmark in data privacy field. We consider Age, Marital Status and Sex from *adult* as the quasi-identifier (QI), and we add a column “Health Condition” as the sensitive attribute, which composes of {HIV, Cancer, Phthisis, Hepatitis, Heart Disease, Asthma, Flu, Indigestion} and its category classification is shown in Table 4.

We randomly choose 400 records from *adult* database as our initial microdata table, and then randomly assign each record in the *adult* a value of the Health Condition. As to the parameters, we choose $\alpha=2$, $p=2$ or 3 and k is 3 or 4. Then, we using p -sensitive algorithm in [15] and switching algorithm in this paper to generate solutions of p -sensitive k -anonymity (Model 1) and (p^+, α) -sensitive k -anonymity models (Model 2) and analyze the number of sensitive attribute disclosures under these two models and compare the

k -anonymity	Number of Health Condition Disclosures	
	2-sensitivity model	$(2^+, 2)$ -sensitivity model
3-anonymity	25	3
4-anonymity	30	4
3-anonymity	3-sensitivity model	$(3^+, 2)$ -sensitivity model
	15	2
4-anonymity	3-sensitivity model	$(3^+, 2)$ -sensitivity model
	21	1

Table 7: Attribute disclosures under two models

running time of these two algorithms (we choose $\alpha=2$, $p=2$ and $k = 2, 4, 6, 8, 10$ to run the experiment).

This experiment shows that even in under enhanced (p^+, α) -sensitive k -anonymity model, disclosure channels still exists so that the Health Condition can be inferred. However, compared with the previous p -sensitive k -anonymity model, our new model could significantly reduce the number of sensitive attribute disclosures. The results of our experiments are summarized in Table 7. As to the running time, there is a large lap among small k . This is because when k is small, the number of QI -groups may be large, so our algorithm has to do a lot of switching process. When k is large, switching algorithm almost performs as good as p -sensitive algorithm (see Figure 1). Moreover, the execution time of switching algorithm increases with α . This is because, when α increases, the number of candidate groups increases, thus the running time increases (see Figure 2).

6. Conclusion

The k -anonymity model protects identity disclosures, but not against attribute disclosures in a microdata table. In this paper, we introduced a new privacy protection model named (p^+, α) -sensitive k -anonymity model. Instead of publishing the specific sensitive values in previous models, we published the categories that the sensitive values belong to. We discussed the properties of this model and presented testing and heuristic algorithms for (p^+, α) -sensitive k -anonymity property. Experimental results showed that our new model can significantly protect the sensitive attributes breach.

7. Acknowledgments

The research is supported by Australian Research Council (ARC) grant DP0774450 titled “*Privacy Preserving Data Sharing in Data Mining Environments*”. Ping Li’s work is supported by the National Natural Science Foundation of China (No. 70501003) and the Aeronautic Science Foundation (No. 2006ZG51079).

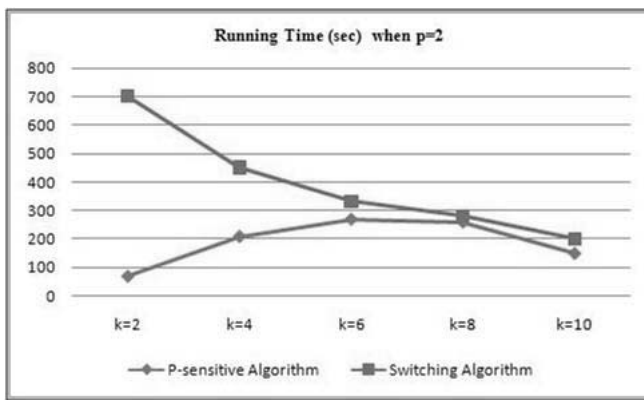


Figure 1: Running time comparison

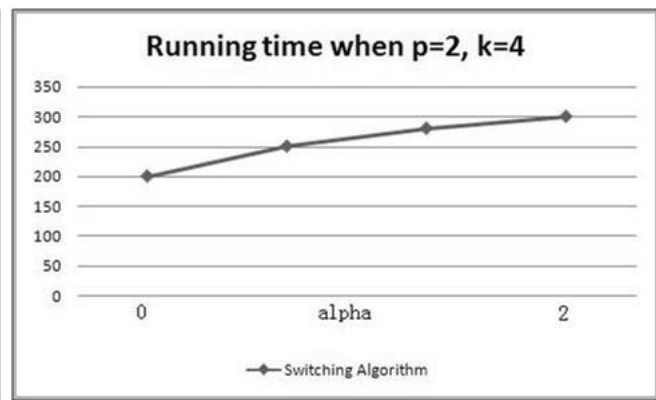


Figure 2: Execution time when α varies

References

- [1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas and A. Zhu. Anonymizing tables. *In Proc. of the 10th International Conference on Database Theory (ICDT05)*, pp. 246-258, Edinburgh, Scotland.
- [2] R. Bayardo and R. Agrawal. Data privacy through optimal k -anonymity. *In Proceedings of the 21st International Conference on Data Engineering (ICDE)*, 2005.
- [3] 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.
- [4] D. Lambert. Measure of disclosure risk and harm. *Journal of Official Statistics*, vol 9, 1993, pp. 313-331.
- [5] K. LeFevre, D. DeWitt and R. Ramakrishnan. Incognito: Efficient Full-Domain k -Anonymity. *In ACM SIGMOD International Conference on Management of Data*, June 2005.
- [6] N. Li, T. Li, S. Venkatasubramanian. t -Closeness: Privacy Beyond k -Anonymity and l -Diversity. *ICDE 2007*: 106-115
- [7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian. l -Diversity: Privacy beyond k -anonymity. *In ICDE*, 2006.
- [8] A. Meyerson and R. Williams. On the complexity of optimal k -anonymity. *In Proc. of the 23rd ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems*, pp. 223-228, Paris, France, 2004.
- [9] D. J. Newman, S. Hettich, C. L. Blake and C. J. Merz. UCI Repository of Machine Learning Databases, available at www.ics.uci.edu/mllearn/MLRepository.html, University of California, Irvine, 1998.
- [10] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010-1027. 2001
- [11] X. Sun, M. Li, H. Wang and A. Plank. An efficient hash-based algorithm for minimal k -anonymity problem. *31st Australasian Computer Science Conference (ACSC 2008)*, Wollongong, NSW, Australia. CRPIT 74, pp: 101-107.
- [12] X. Sun, H. Wang and J. Li. On the complexity of restricted k -anonymity problem. *10th Asia Pacific Web Conference (APWEB2008)*, LNCS 4976, pp: 287-296, Shenyang, China.
- [13] 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.
- [14] L. Sweeney. k -Anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty Fuzziness Knowledge-based Systems*, 10(5), pp 557-570, 2002
- [15] 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.
- [16] T. M. Truta, A. Campan and P. Meyer. Generating Microdata with P -sensitive k -anonymity Property. *SDM 2007*: 124-141
- [17] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up Generalization: A Data Mining Solution to Privacy Protection. *The fourth IEEE International Conference on Data Mining (ICDM2004)* 249-256.
- [18] W. E. Winkler. Advanced Methods for Record Linkage, *Proceedings of the Section on Survey Research Methods, American Statistical Society*, 467-472
- [19] 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.