STACK AND DEAL: AN EFFICIENT ALGORITHM FOR PRIVACY PRESERVING DATA PUBLISHING

Vikas Thammanna Gowda

Department of Electrical Engineering and Computer Science,
Wichita State University, Kansas, USA

ABSTRACT

Although k-Anonymity is a good way to publish microdata for research purposes, it still suffers from various attacks. Hence, many refinements of k-Anonymity have been proposed such as l-diversity and t-Closeness, with t-Closeness being one of the strictest privacy models. Satisfying t-Closeness for a lower value of t may yield equivalence classes with high number of records which results in a greater information loss. For a higher value of t, equivalence classes are still prone to homogeneity, skewness, and similarity attacks. This is because equivalence classes can be formed with fewer distinct sensitive attribute values and still satisfy the constraint t. In this paper, we introduce a new algorithm that overcomes the limitations of k-Anonymity and l-Diversity and yields equivalence classes of size k with greater diversity and frequency of a SA value in all the equivalence classes differ by at-most one.

KEYWORDS

k-Anonymity, l-Diversity, t-Closeness, Privacy Preserving Data Publishing.

1. INTRODUCTION

Various organizations such as government agencies and hospitals release microdata for medical research, trend analysis, and other purposes. Typically, microdata is stored in a table and each row corresponds to an individual's record and each record consists of a diverse number of attributes. These attributes can be categorized into a)*Explicit Identifier* attributes: are attribute sets such as name and social security number, that explicitly identify individuals. b) *Quasi Identifier* (QI) attributes: are attribute sets such as zip code, age, and sex that cannot uniquely identify individuals, but combinations of these attributes can give away the record holder. Sweeney [1] has shown that even though neither sex, date of birth, nor zip codes uniquely identifies an individual, the combination of all three is sufficient to identify 87% of individuals in the United States. c) *Sensitive attributes* (SAs): consists of sensitive information of individuals. d) *Non-Sensitive attributes*: consists of attributes that are non-sensitive in nature which does notreveal any sort of information about the record holder.

Privacy preserving data publishing (PPDP) means releasing microdata in such a way that there is data utility of released data and at the same time privacy of an individual in the released data is maintained. Prior to data release, first, the explicit identifier attributes are removed since it uniquely identifies an individual. Then the records are horizontally partitioned into groups of records called equivalence classes and the quasi identifier attributes are generalized to ensure that quasi identifier values of all records within an equivalence class becomes identical while the sensitive attributes are unaltered.

David C. Wyld et al. (Eds): CMC, NCO, SOFT, CDKP, MLT, ICAITA - 2021
pp. 139-148, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111111

Based on this approach, various privacy models have been proposed. For example, *k*-anonymity (Sweeney [1]) requires that each equivalence class must have at least *k* records that are indistinguishable from *k-1* records in terms of their quasi identifier attribute values. *l*-diversity (Machanavajjhala et al. [2]) requires that each equivalence class consists of at least a certain number of i.e., *l* "*well-represented*" values of sensitive attributes. To address the limitations of *k*-anonymity and *l*-diversity Li et al. [3] introduced the concept of *t*-closeness [9], which requires that distance between the distribution of the sensitive attribute in the entire table and the distribution of the sensitive attribute in any equivalence class to be close.

l-diversity and *t*-closeness privacy models are the extensions of *k*-anonymity model to address its limitations. This paper shows that the limitations can be addressed with an algorithm since the extensions possess its own limitations. The algorithm outputs equivalence classes with a high degree of diversity among the sensitive attributes whose distribution is very close to the distribution of sensitive attributes in the overall table with just one input parameter *k*. The algorithm can be implemented with the help of simple data structures like queue or stack.

1.1. Contributions and Organization

In this paper, we have introduced an algorithm which gives equivalence classes whose sensitive attribute distribution is close to sensitive attribute distribution in the overall table and overcomes the limitations of *k*-Anonymity and *l*-Diversity. The rest of the paper is organized as follows. In Section 2, we review some background concepts used throughout the paper. Section 3 deals with our proposed method that works in various stages and provides the algorithm for obtaining equivalence classes of size *k* with greater diversity and frequency of a SA value in all the ECs differ by at-most one. In Section 4, we analyse the algorithm and show how it defends against homogeneity, skewness and similarity attacks with experimental results and Section 5 presents conclusion and future work.

2. BACKGROUND

Consider a raw data that needs to be published as shown in Table 1. Explicit identifiers such as name and SSN are removed since they directly identify the record holder. Quasi identifiers like zip code and age cannot uniquely identify individuals but, combinations of these attributes can give away the record holder. Sweeney [1] has shown that even though neither sex, date of birth nor zip codes uniquely identify an individual, the combination of all three is sufficient to identify 87% of individuals in the United States. Attribute like disease that is closely guarded by the record holder is considered to be sensitive attribute.

Table 1. Raw Table.

No	Name	SSN	Zip Code	Age	Disease
1	Scofield	111-11-1111	47677	29	Flu
2	Linc	222-22-2222	47602	25	Flu
3	Sara	333-33-3333	47678	27	Flu
4	Henry	444-44-4444	47905	43	Cancer
5	Bagwell	555-55-5555	47909	40	Ulcer
6	Bellick	666-66-6666	47706	47	Cold
7	John	777-77-7777	47705	30	Cancer
8	Cooper	888-88-8888	47773	35	Pneumonia
9	Sucre	999-99-9999	47707	32	Bronchitis

The goal of PPDP is to protect the sensitive attribute of the record holder while still publishing enough information to maintain data utility. *k*-anonymity by Sweeney [1] is a well-known model for anonymizing the data. Here the explicit identifiers of each record are removed and quasi identifiers along with sensitive attribute are grouped. Each group is called an equivalence class where quasi identifiers are generalized and sensitive attribute is unaltered.

Definition 1: (*Equivalence Class*) An Equivalence Class is a set of anonymized records that have same values for all quasi identifier attributes, i.e., all records in each equivalence class are indistinguishable in terms of their quasi identifier attributes.

Definition 2: (*k-Anonymity*) An equivalence class is said to satisfy *k*-anonymity if every record is indistinguishable from at least *k-1* other records with respect to every set of the quasi identifier attributes. A table is said to satisfy *k*-anonymity if every equivalence class of the table satisfies *k*-anonymity.

In other words, it is like hiding something in the crowd so it would be difficult to identify, as almost everything looks alike when the entire crowd is seen.

Table 2 gives a 3-anonymous version of the raw table. The data is divided into three equivalence classes consisting of three records each, whose quasi identifiers (zip code and age) are generalized and sensitive attribute (disease) is unaltered.

Table 2. 3-Anonymous Version of Table 1.

No	Zip Code	Age	Disease
1	476**	2*	Flu
2	476**	2*	Flu
3	476**	2*	Flu
4	479**	4*	Cancer
5	479**	4*	Ulcer
6	479**	4*	Cold
7	477**	3*	Cancer
8	477**	3*	Pneumonia
9	477**	3*	Bronchitis

Attack on *k*-Anonymity: Suppose that Alex and Bob are neighbours and Alex discovers a published data as shown in Table 2. Alex knows that Bob is a 29-year old male living in zip code 47677, then Alex can easily place Bob in first equivalence class. Since all the record holders in first equivalence class of Table 2 have the same disease i.e., flu, Alex concludes that Bob has flu. This is known as homogeneity attack.

Limitations of *k*-Anonymity:

1. Does not provide protection against homogeneity attack.
2. Does not include randomization and attacker can still make inferences about data sets that may harm individuals.
3. Not good for high dimensional data.
4. Concerned only about quasi identifiers and not sensitive attribute.

Machanavajjhala et al. [2] introduced *l*-diversity as a stronger notion of privacy to overcome the limitations of k-anonymity.

Definition 3: (*l-Diversity*) An equivalence class is said to satisfy *l*-diversity if there are at-least *l* "well represented" values for the sensitive attribute. A table is said to satisfy *l*-diversity if every equivalence class of the table satisfies *l*-diversity.

Table 4 satisfies 3-diversity since there are three well represented sensitive attribute values in each equivalence class. The table also satisfies 3-anonymity.

Attack on *l*-Diversity: Suppose that Alex and Bob are neighbours and Alex discovers a published data as shown in Table 4. Alex knows that Bob is a 37-year old male living in zip code 67220, then Alex can easily place Bob in first equivalence class. Looking at the SA values, Alex concludes that Bob is suffering from some sort of stomach related disease. This is known as similarity attack. *l*-diversity fails to protect against attacks arising from an adversary's unavoidable knowledge of the overall distribution of SA values in a released table. A skewness attack may occur when the distribution of sensitive attributes in an equivalence varies significantly from that in the released table.

Table 3. Disease Table.

No	Zip Code	Age	Disease
1	67200	37	Gastric ulcer
2	67406	52	Gastritis
3	67207	35	Gastritis
4	67433	57	Flu
5	67319	41	Bronchitis
6	67302	43	Pneumonia
7	67308	46	Stomach cancer
8	67420	58	Bronchitis
9	67208	36	Stomach cancer

Table 4. 3-Diverse Version of Table 3.

No	Zip Code	Age	Disease
1	672**	3*	Gastric ulcer
2	672**	3*	Gastritis
3	672**	3*	Stomach cancer
4	674**	5*	Gastritis
5	674**	5*	Flu
6	674**	5*	Bronchitis
7	673**	4*	Bronchitis
8	673**	4*	Pneumonia
9	673**	4*	Stomach cancer

Limitations of *l*-Diversity:

1. Does not provide protection against similarity and skewness attacks.
2. *l*-diversity may be difficult and unnecessary to achieve.
3. It is concerned only about well represented sensitive attributes but not about the distribution of the sensitive attributes.

4. ALGORITHM FRAMEWORK

In this section, we present a framework for Stack and Deal algorithm. Given a microdata table*M* consisting of *r* records and *n* attributes ((*n-1*) quasi identifier attributes and one sensitive attribute) and *k*, let *A* denote the set of all attributes $\{A_1, A_2, ..., A_n\}$. Without loss of generality, let the attribute A_n be the sensitive attribute and $\{A_1, A_2, ..., A_{n-1}\}$ be quasi identifier attributes.

Stage 1: Frequency and Distribution of SA in the entire table M

A frequency table as shown in Table 5 is created that contains *s* sensitive attribute values ($S_1, S_2, S_3, ..., S_s$) and its frequency $F = (f_1, f_2, f_3, ..., f_s)$ in the entire table.

Table 5. Frequency Distribution Table of Sensitive attribute in M.

No	Sensitive Attribute	Frequency	Distribution
1	S_1	f_1	p_1
2	S_2	f_2	p_2
3	S_3	f_3	p_3
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
s	S_s	f_s	p_s

These entries are arranged in descending order, where $(f_1 \geq f_2 \geq f_3 \geq ..... \geq f_s)$ and $\sum_{j=1}^{s} f_j = r$. Distribution of the sensitive attribute in the entire table is $P = (p_1, p_2, p_3, ..., p_s)$, where $(p_1 \geq p_2 \geq p_3 \geq ..... \geq p_s)$, $p_j = f_j / r$ and $\sum_{j=1}^{s} p_j = 1$.

Stage 2: Stack and Deal the records

In this stage, a queue of records are stacked according to the frequency distribution table as shown in Table 6 i.e., all records having sensitive attribute value S_1 appears at the top of the queue and records having sensitive attribute value S_s appears at the bottom of the queue.

Table 6. Stacked Data.

No	Quasi Identifier	Sensitive Attribute
1	$\{A_1, A_2, A_{3,....,} A_{n-1}\}$	S_1
2		S_1
.		S_1
.		.
.		.
33		S_2
.		.
.		.
.		.
87		S_{s-1}
.		.
.		.
.		.
r	$\{A_1, A_2, A_{3,....,} A_{n-1}\}$	S_s

Now for dealing part, each record is popped out of the stack into *e* equivalence classes (*e = r/k)* in a cyclic order. For example, if there are ten equivalence classes then, the first record goes into first equivalence class, second record to second equivalence class and so on. When we hit the last equivalence class i.e., tenth equivalence the next record goes into the first equivalence class and the cycle continues till the stack is empty.

Observation: We see that, by following the cyclic order while populating equivalence classes we get equi-sized equivalence classes where every equivalence will get equal portions of f_j/e and frequency of a SA value in all the equivalence classes differs by at-most one.

Stage 3: Frequency and Distribution of SA in equivalence classes *E*

Once the last record is popped out, we now have *e* equivalence classes, $E = (E_1, E_2, E_3, \ldots, E_e)$ having *k* records. Similar to stage 1, frequency and distribution of SA in each equivalence class is formed, that contains sensitive attribute values $(S_1, S_2, S_3, \ldots, S_s)$ and its frequency $F = (g_1, g_2, g_3, \ldots, g_s)$. These entries are arranged in descending order, where $(g_1 \geq g_2 \geq g_3 \geq \ldots\ldots \geq g_s)$ and $\sum_{j=1}^{s} g_j = k$. Distribution of the sensitive attribute in an equivalence class is $Q = (q_1, q_2, q_3, \ldots, q_s)$, where $(q_1 \geq q_2 \geq q_3 \geq \ldots\ldots \geq q_s)$, $q_j = g_j / r$ and $\sum_{j=1}^{s} q_j = 1$. Distribution table of one equivalence class is shown below in Table 7. Earth movers distance [8] between *P* and *Q* gives the closeness between SA distribution in the overall table and the SA distribution in each equivalence class.

Table 7. Frequency Distribution Table of Sensitive attribute in an Equivalence Class.

No	Sensitive Attribute	Frequency	Distribution
1	S_1	g_1	q_1
2	S_2	g_2	q_2
3	S_3	g_3	q_3
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
s	S_s	g_s	q_s

Algorithm:

Input: micro table *M* having *r* records, *k*

Output: *e* equivalence class of size *k*

1. Let $e = r/k$.
2. Set $E_1, E_2, E_3, \ldots, E_e = \phi$
3. Sort all records in descending order of f_j (frequency of SA $(1 \geq j \geq s)$)
4. For $z = 1$ to r

$E[(z \bmod e) + 1] = M[z]$

5. ANALYSIS OF ALGORITHM FOR VARIOUS ATTACKS

In this section, we show how the Stack and Deal algorithm protects against various attacks:

Protection against Homogeneity attack:

Homogeneity attack occurs when the SA values in an EC are the same, thus an attacker learns about the sensitive information of a record holder without any additional efforts. The way to combat this is to ensure that the SA values in every EC are diverse. Our algorithm ensures that all the ECs produced are diverse in terms of their SA values.

Let F = (167, 153, 127, 103, 91, 89) and r =730. When we vary the value of k we observe that we attain maximum diversity for k = 9. We know that if an EC satisfy 9-anonymity it also satisfies 2, 3, ..., 8-anonymity as well. Since there is a trade-off between privacy and data utility, we can compromise data utility to achieve maximum diversity. Figure 1 shows the variation of k with respect to l.

We run the same experiment on Adult data set Figure 2 from UC Irvine machine learning repository and vary k from 2 to 21. We observe relatively similar behaviour on this data set too.

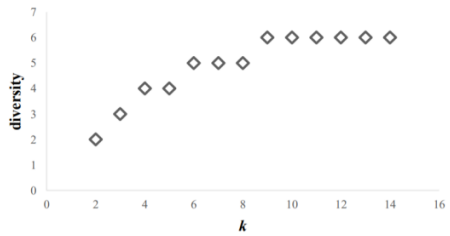

Figure 1. k vs diversity

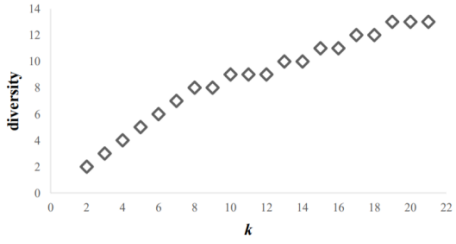

Figure 2. k vs diversity for Adult Data Set.

Protection against Skewness and Similarity attacks:

Privacy is measured by the amount of information gain of an observer/attacker. The observer has some prior belief (G_0) about the sensitive information of a record holder and some posterior belief (G_2) after seeing the released table. Information gain is the difference in these two believes. Assume that the observer is given a completely generalized form of the data P and his prior belief (G_0) changes to (G_1) by looking at the distribution of SA values in the overall table P (P is considered as public information because as long as a version of data is released, P will be known). Now, the observer is given the released data and by knowing the quasi-identifier of a record holder, the observer is able to identify an EC to which the record holder belongs to and learns the distribution of SA values represented as Q in that EC. Now this is the observer's posterior belief (G_2).

The l-diversity requirement is inspired by restricting the difference between prior belief and observer's posterior belief but, whenever the distribution of SA values within an EC varies significantly from their overall distribution in the released table. l-diversity fails to guarantee privacy allowing skewness and similarity attacks. In our method, we choose to limit the difference between (G_1) and (G_2). We can do this by ensuring that the frequencies of SA values in all the ECs are similar and limiting their difference to be as low as possible. This is because, we want to obtain ECs which are of equi sized so as to limit the information loss and if the difference in frequencies increases, Q moves further away from P. Thus, by limiting the difference in the frequencies of SA values in the EC, we can limit the difference between P and Q and there by finally limiting the gain from (G_1) to (G_2). The distance between these two distributions is calculated using earth movers distance [8].

Earth Movers Distance: For any two distributions P and Q, where $P = (p_1, p_2, p_3, \ldots, p_s)$, $Q = (q_1, q_2, q_3, \ldots, q_s)$ and $\sum_{i=1}^{s} p_i = \sum_{i=1}^{s} q_i = 1$, the earth movers distance between P and Q, denoted as $EMD\ (P,Q)$.

$$EMD\ (P,Q) = \frac{1}{s-1}\sum_{i=1}^{s}\sum_{j=1}^{i}|(p_j - q_j)|$$

The earth mover's distance can be thought of as the sum total of the portions of the p_i values that needs to be moved to other indices in P each portion scaled by the normalized distance of its movement within the m-tuple, to turn P into Q.

As an example, consider probability distributions,

P = (0.2, 0.1, 0.7)
Q = (0.3, 0.0, 0.7)
R = (0.1, 0.0, 0.9)

$EMD\ (P, Q)$ = 0.1(1/2) = 0.05, because in order to turn P into Q, 0.1 amount needs to be moved from p_2 to p_1, which is 1 index away, out of a maximum of 2 (as k-1 = 2 is the farthest movement distance in this tuple). Similarly, $EMD\ (Q, R)$ = 0.2(2/2) = 0.2 and $EMD\ (P, R)$ = 0.1(2/2) + 0.1(1/2) = 0.15.

To study the result, we plot k against EMD between P and Q of ECs generated using our algorithm and randomly generated ECs. We observe that difference between $P$and Q reduces as we increase k and our algorithm gives the minimum difference. Figure 3 represents the plot for F = (167, 153, 127, 103, 91, 89) and r=730 and varying k. We observe that for k=2 we get some ECs whose difference between P and Q is lesser than our algorithm, this is because the size of ECs for such values vary by a huge difference increasing the information loss.

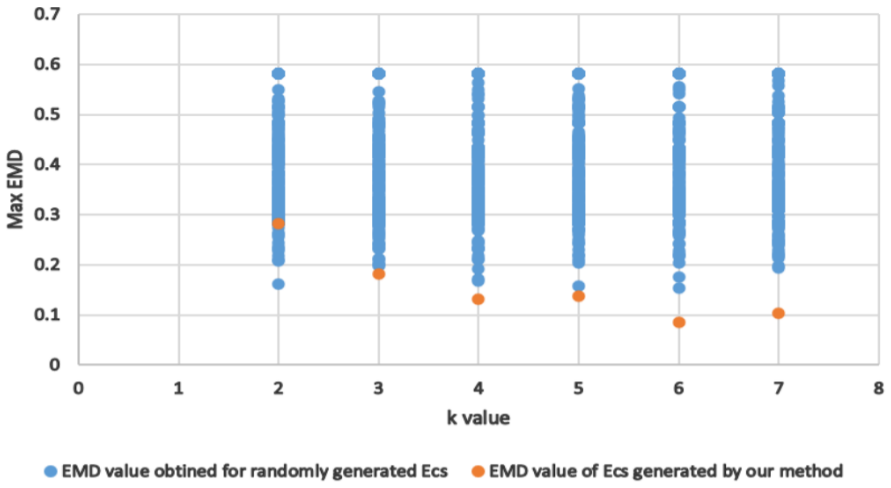

Figure 3. k vs EMD

Next, let us study the effect of increasing the difference between SA values in the ECs. For this purpose, we use Blood Transfusion data set and Haberman's Survival data set from UC Irvine machine learning repository and vary k from 2 to 20. *Rand1* and *Rand2* are the set of ECs whose difference in frequency of SA values are 2 and 3, respectively. From Figure 4 and Figure 5 we

observe that by limiting the difference in the frequencies of SA values in the EC we can limit the difference between P and Q and thereby finally limiting the gain from G_1 to G_2.

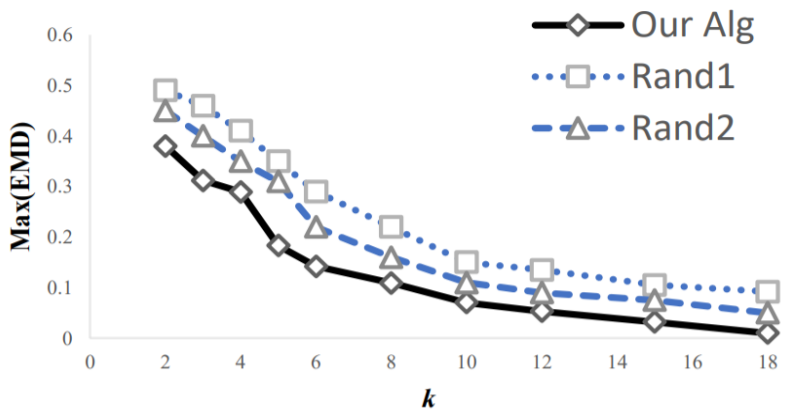

Figure 4. Variation of k in Blood Transfusion data set.

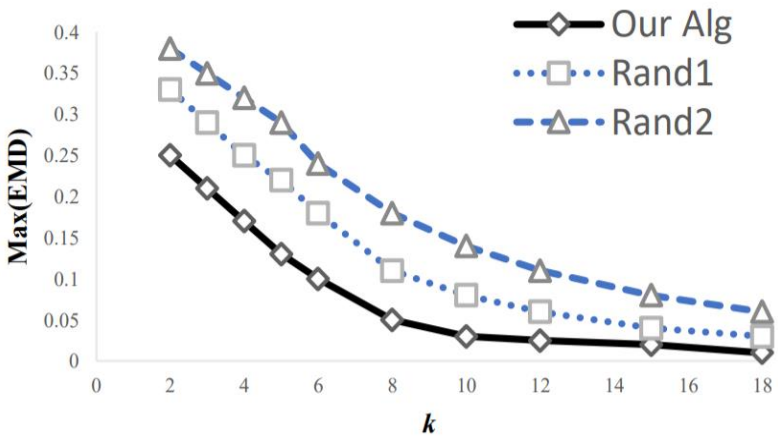

Figure 5. Variation of *k* in Haberman's Survival data set.

6. CONCLUSION AND FUTURE WORK

While *k*-Anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. *l*-Diversity seeks to solve this problem by adding a condition that each equivalence class must have *l* distinct SA values. We have seen the limitations of *l*-Diversity and how we can combat them with the help of our algorithm without the requirement of *t* in *t*-Closeness. We have introduced a new algorithm that takes the input parameter *k* along with the microdata and produces equivalence classes of size *k* with a greater diversity and frequency of a SA value in all the ECs differ by at-most one thus helping in minimal data loss.

The first direction of future work is to design an algorithm that exchanges records to minimize information loss till we reach an optimal value for the information loss by making use of the parameter *t.* As a second direction, this algorithm can be generalized for Multiple Sensitive Attributes.

REFERENCES

[1] Latanya Sweeney.: k-anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5), pages 557-570, 2002.

[2] Machanavajjhala, Ashwin, Gehrke, Johannes, Kifer, Daniel, Venkitasubramaniam, Muthuramakrishnan: l-Diversity: Privacy Beyond k-Anonymity ACM Transactions on Knowledge Discovery From Data - TKDD. 1. 24. 10.1145/1217299.1217300.

[3] L. Ninghui, L. Tiancheng, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity", Proc.-Int. Conf. Data Eng., no. 3, pp. 106-115, 2007

[4] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain K- anonymity," in Proceedings of the ACM SIGMOD International Conference on 77 Management of Data, 2005, vol. 10, no. 5, pp. 49–60.

[5] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," in 22nd International Conference on Data Engineering (ICDE'06), 2006, vol. 2006, pp. 25–25

[6] J. Cao, P. Karras, P. Kalnis, and K.-L. Tan, "SABRE: a Sensitive Attribute Bucketization and REdistribution framework for t-closeness," VLDB J., vol. 20, no. 1, pp. 59–81, Feb. 2011.

[7] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, "t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation," IEEE Trans. Knowl. Data Eng., vol. 27, no. 11, pp. 3098–3110, Nov. 2015

[8] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," International Journal of Computer Vision, vol. 40, no. 2, pp. 99–121, 2000.

[9] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. Closeness: A new privacy measure for data publishing. IEEE Trans. Knowledge and Data Engineering, 22(7), 2010.

[10] Tiancheng Li , Jian Zhang , Ian Molloy , "Slicing: A New Approach for Privacy Preserving Data Publishing" IEEE Transaction on KDD (2012).

[11] B.Vani, D.Jayanthi, "Efficient Approach for Privacy Preserving Microdata Publishing Using Slicing" IJRCTT 2013.

[12] S.Gokila, Dr.P.Venkateswari, A SURVEY ON PRIVACY PRESERVING DATA PUBLISHING International Journal on Cybernetics & Informatics (IJCI) Vol. 3, No. 1, February 2014

[13] M. Patel, P. Richariya, and A. Shrivastava. A review paper on privacy preserving data mining. Compusoft, 2(9):296, 2013.

[14] C. C. Aggarwal and P. S. Yu. A general survey of privacy-preserving data mining models and algorithms. Privacy-preserving data mining, pages 11-52, 2008.

[15] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3):283-304, 1998.

[16] J. Byun, A. Kamra, E. Bertino, and N. Li. Efficient k-anonymization using clustering techniques. pages 188-200, 2007.

[17] Q. Wei, Y. Lu, and Q. Lou. Privacy-preserving data publishing based on declustering. Pages 152-157, 2008.

[18] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. pages 139-150, 2006.

[19] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. Pages 116-125, 2007.

[20] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent development. ACM Comput. Surv., 42(4):14:1-14:53, June 2010.

[21] C. Dwork. Differential Privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP). 1–12, Venice, Itily, July 2006.