A Privacy Measure for Data Disclosure to Publish Micro Data using (n, t)-Closeness

A. Sunitha
Assistant Professor, Dept of IT
St.Martin's Engineering College
Secunderabad, A.P, India

K. Venkata Subba Reddy
Assistant Professor, Dept of CSE
Muffakham Jah College of
Engineering & Technology
Hyderabad, A.P, India

B. Vijayakumar
Professor & HOD, Dept of IT
St.Martin's Engineering College
Secunderabad, A.P, India

ABSTRACT
Closeness is described as a privacy measure and its advantages are illustrated through examples and experiments on a real dataset. In this Paper the closeness can be verified by giving different values for N and T. Government agencies and other organizations often need to publish micro data, e.g., medical data or census data, for research and other purposes. Typically, such data are stored in a table, and each record (row) corresponds to one individual. Generally if we want to publish micro data A common anonymization approach is generalization, which replaces quasi-identifier values with values that are less-specific but semantically consistent. As a result, more records will have the same set of quasi-identifier values. An equivalence class of an anonymized table is defined to be a set of records that have the same values for the quasi-identifiers
To effectively limit disclosure, the disclosure risk of an anonymized table is to be measured. To this end, k-anonymity is introduced as the property that each record is indistinguishable with at least k-1 other records with respect to the quasi-identifier i.e., k-anonymity requires that each equivalence class contains at least k records. While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. To address the above limitation of k-anonymity, a new notion of privacy, called l-diversity is introduced, which requires that the distribution of a sensitive attribute in each equivalence class has at least l "well represented" values. One problem with l-diversity is that it is limited in its assumption of adversarial knowledge. This assumption generalizes the specific background and homogeneity attacks used to motivate l-diversity. The k-anonymity privacy requirement for publishing micro data requires that each equivalence class contains at least k records. But k-anonymity cannot prevent attribute disclosure. The notion of l-diversity has been proposed to address this; l-diversity requires that each equivalence class has at least l well-represented values for each sensitive attribute. L-diversity has a number of limitations. In particular, it is neither necessary nor sufficient to prevent attribute disclosure. Due to these limitations, a new notion of privacy called "closeness" is proposed. First the base model t- closeness is presented, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. Then a more flexible privacy model called (n, t)-closeness is proposed. The rationale for using

General Terms
Data and Knowledge based systems. Privacy Preservation

Keywords
Privacy Measure,K-Anonomity,L-Diversity, data Anonymization,(n-t) closeness

.

1. INTRODUCTION
Government agencies and other organizations often need to publish micro data, e.g., medical data or census data, for research and other purposes. Typically, such data are stored in a table, and each record (row) corresponds to one individual.
Each record has a number of attributes, which can be divided into the following three categories:
1) Attributes that clearly identify individuals. These are known as explicit identifiers and include, e.g., Social Security Number.
2) Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers, and may include, e.g., Zip code, Birth-date and Gender.
3) Attributes that are considered sensitive, such as Disease and Salary. While the released table gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the table.
Therefore, the objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonymizing the data before release. The first step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly available databases (e.g., a voter registration list) that include both explicit identifiers and quasi –identifiers
A common anonymization approach is generalization, which replaces quasi-identifier values with values that are less-specific but semantically consistent. As a result, more records will have the same set of quasi-identifier values. An equivalence class of an anonymized table is defined to be a set of records that have the same values for the quasi-identifiers
To effectively limit disclosure, the disclosure risk of an anonymized table is to be measured. To this end, k-anonymity is introduced as the property that each record is indistinguishable with at least k-1 other records with respect to the quasi-identifier i.e., k-anonymity requires that each equivalence class contains at least k records. While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure.
To address the above limitation of *k*-anonymity, a new notion of privacy, called *l*-diversity is introduced, which requires that the distribution of a sensitive attribute in each equivalence class has at least *l* "well represented" values. One problem with *l*-diversity is that it is limited in its assumption of adversarial knowledge. This assumption generalizes the specific background and homogeneity attacks used to motivate *l*-diversity. Therefore, the proposed privacy measure

limits the amount of individual-specific information an observer can learn.

2. PROBLEM DEFINITION

The objective of this work is to limit the disclosure risk to an acceptable level while maximizing the benefit. While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. The notion of '*l*-diversity attempts to solve this problem. *L-diversity* presents two attacks. Due to these limitations, a novel privacy notion called "closeness" is proposed .Two instantiations are proposed: a base model called t-closeness and a more flexible privacy model called (n, t)-closeness.

3. *K*-ANONIMITY

Definition 1. *K-anonymity*

Let RT (*A1*... *An*) be a released table and QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy *k*-anonymity if and only if each sequence of values in T[QI_{RT}] appears with at least *k* occurrences in RT[QI_{RT}].

Example 1. Table adhering to *k*-anonymity

Figure 1 provides an example of a table T that adheres to *k*-anonymity. The quasi-identifier for the table is QI_T= {*Race*, *Birth*, *Gender*, *ZIP*} and *k*=2. Therefore, for each of the tuples contained in the table T, the values of the tuple that comprise the quasi-identifier appear at least twice in T. That is, for each sequence of values in T[QI_T] there are at least *2* occurrences of those values in T[QI_T]. In particular, *t1*[QI_T] = *t2*[QI_T], *t3*[QI_T] = *t4*[QI_T], *t5*[QI_T] = *t6*[QI_T], *t7*[QI_T] = *t8*[QI_T] = *t9*[QI_T], and *t10*[QI_T] = *t11*[QI_T].

Table1: Example of *K*-anonymity

	Non-Sensitive			Sensitive
	Zip Code	**Age**	**Nationality**	**Condition**
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Example 2. *K* occurrences of each value under *k*-anonymity

Table T in Figure 1 adheres to *k*-anonymity, where QI_T = {*Race*, *Birth*, *Gender*, *ZIP*} and *k*=2. Therefore, each value that appears in a value associated with an attribute of QI in T appears at least *k* times. |T[*Race* ="black"]| = 6. |T[*Race* ="white"]| = 5. |T[*Birth* ="1964"]| = 5. |T[*Birth* ="1965"]| = 4. |T[*Birth* ="1967"]| = 2. |T[*Gender* ="m"]| = 6. |T[*Gender* ="f"]| = 5. |T[*ZIP* ="0213*"]| = 9. And, |T[*ZIP* ="0214*"]| = 2 [3]. Table 3 4-anonymous Inpatient Micro data

A table satisfies *k*-anonymity if every record in the table is indistinguishable from at least $k-1$ other records with respect to every set of quasi-identifier attributes; such a table is called a *k-anonymous* table. Hence, for every combination of values of the quasi-identifiers in the *k*-anonymous table, there are at least *k* records that share those values. This ensures that individuals cannot be uniquely identified by linking attacks.

An attribute is marked *sensitive* if an adversary must not be allowed to discover the value of that attribute for any individual in the dataset. Attributes not marked sensitive are *non-sensitive*. Let the collection of attributes {zip code, age, nationality}

Let the collection of attributes {zip code, age, nationality} be the quasi-identifier for this dataset. Figure 3 shows a 4-anonymous table derived from the table in Figure 2 (here "*" denotes a suppressed value so, for example, "zip code = 1485*" means that the zip code is in the range [14850−14859] and "age=3*"means the age is in the range [30 − 39]). Note that in the 4-anonymous table, each tuple has the same values for the quasi-identifier as at least three other tuples in the table

Table 2: Inpatient Micro data

Race	Birth	Gender	ZIP	Problem
Black	1965	M	0214*	Short breath
Black	1965	M	0214*	Chest pain
Black	1965	F	0213*	Hypertension
Black	1965	F	0213*	Hypertension
Black	1964	F	0213*	Obesity
Black	1964	F	0213*	Chest pain
White	1964	M	0213*	Chest pain
White	1964	M	0213*	Obesity
White	1964	M	0213*	Short breath
White	1967	M	0213*	Chest pain
White	1967	M	0213*	Chest pain

Let the collection of attributes {zip code, age, nationality} be the quasi-identifier for this dataset. Figure 3 shows a 4-anonymous table derived from the table in Figure 2 (here "*" denotes a suppressed value so, for example, "zip code = 1485*" means that the zip code is in the range [14850−14859] and "age=3*"means the age is in the range [30 − 39]). Note that in the 4-anonymous table, each tuple has the same values for the quasi-identifier as at least three other tuples in the table.

3.1. Attacks on k-Anonymity

This section presents two attacks, the *homogeneity attack* and the *background knowledge attack*, and shows how they can be used to compromise a k-anonymous dataset. Homogeneity Attack: Alice and Bob are antagonistic neighbors. One day Bob falls ill and is taken by ambulance to the hospital. Having seen the ambulance, Alice sets out to discover what disease Bob is suffering from. Alice discovers the 4-anonymous table of current inpatient records published by the hospital (Figure 3), and so she knows that one of the records in this table contains Bob's data. Since Alice is Bob's neighbor, she knows that Bob is a 31-year-old American male who lives in the zip code 13053. Therefore, Alice knows that Bob's record number is 9, 10, 11, or 12.Now, all of those patients have the same medical condition (cancer), and so Alice concludes that Bob has cancer.

Table 3: 4-Anonymous Inpatient Micro data

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	1485*	>_40	*	Cancer
6	1485*	>_40	*	Heart Disease
7	1485*	>_40	*	Viral Infection
8	1485*	>_40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Observation 1: k-Anonymity can create groups that leak information due to lack of diversity in the sensitive attribute. Note that such a situation is not uncommon. As a back-of-the-envelope calculation, suppose we have a dataset containing 60,000 distinct tuples where the sensitive attribute can take 3 distinct values and is not correlated with the no sensitive attributes. A 5-anonymization of this table will have around 12,000 groups2 and, on average, 1 out of every 81 groups will have no diversity (the values for the sensitive attribute will all be the same). Thus we should expect about 148 groups with no diversity. Therefore, information about 740 people would be compromised by a homogeneity attack.

This suggests that in addition to k-anonymity, the sanitized table should also ensure "diversity" – all tuples that share the same values of their quasi-identifiers should have diverse values for their sensitive attributes. Background Knowledge Attack: Alice has a pen friend named Umeko who is admitted to the same hospital as Bob, and whose patient records also appear in the table shown in Figure 3. Alice knows that Umeko is a 21 year old Japanese female who currently lives in zip code 13068. Based on this information, Alice learns that Umeko's information is contained in record number 1, 2, 3, or 4. Without additional information, Alice is not sure whether Umeko caught a virus or has heart disease. However, it is well known that Japanese have an extremely low incidence of heart disease. Therefore Alice concludes with near certainty that Umeko has a viral infection.

Observation 2: k-Anonymity does not protect against attacks based on background knowledge [4]

4. *l*-DIVERSITY

To address these limitations of k-anonymity, l-diversity is introduced as a stronger notion of privacy. The l-diversity principle: An equivalence class is said to have l-diversity if there are at least l "well-represented" values for the sensitive attribute. A table is said to have l- diversity if every equivalence class of the table has l-diversity.

Following are a number of interpretations of the term "well represented" in this principle:

1. Distinct l-diversity. The simplest understanding of "well represented" would be to ensure that there are at least l distinct values for the sensitive attribute in each equivalence class. Distinct l-diversity does not prevent probabilistic inference attacks. An equivalence class may have one value appear much more frequently than other values, enabling an adversary to conclude that an entity in the equivalence class is very likely to have that value. This motivated the development of the following stronger notions of l-diversity.

2. Entropy l-diversity. The entropy of an equivalence class E is defined to be

$$\text{Entropy (E)} = -\sum_{s \in S} p(E,s) \log p(E,s)$$

in which S is the domain of the sensitive attribute and p(E, s) is the fraction of records in E that have sensitive value s. A table is said to have entropy l-diversity if for every equivalence class E, Entropy (E) ≥ log l. Entropy l-diversity is stronger than distinct l-diversity. As pointed out in [4], in order to have entropy l- diversity for each equivalence class, the entropy of the entire table must be at least log (l). Sometimes, this may too restrictive, as the entropy of the entire table may be low if a few values are very common. This leads to the following less conservative notion of l-diversity.

3. Recursive (c, l)-diversity**.** Recursive (c, l)-diversity (c is a float number and l is an integer) makes sure that the most frequent value does not appear too frequently, and the less frequent values do not appear too rarely. Let m be the number of values in an equivalence class, and r_i, $1 \le i \le m$ be the number of times that the i th most frequent sensitive value appears in an equivalence class E.Then, E is said to have recursive (c, l)-diversity if $r1 < c (r_l + r_{l+1} + \ldots + r_m)$. A table is said to have recursive (c, l) -diversity if all of its equivalence classes have recursive (c, l)-diversity [5].

Limitations of l -Diversity

While the l-diversity principle represents an important step beyond k-anonymity in protecting against attribute disclosure, it has several shortcomings shown below l-diversity is insufficient to prevent attribute disclosure. Two attacks on l-diversity are shown below.

Skew ness Attack: When the overall distribution is skewed, satisfying l-diversity does not prevent attribute disclosure. Consider again Example 5. Suppose that one equivalence class has an equal number of positive records and negative records. It satisfies distinct 2-diversity, entropy 2-diversity, and any recursive (c, 2)-diversity requirement that can be imposed. However, this presents a serious privacy risk, because anyone in the class would be considered to have 50% possibility of being positive, as compared with the 1% of the overall population.

Now consider an equivalence class that has 49 positive records and only 1 negative record. It would be distinct 2-diverse and has higher entropy than the overall table (and thus satisfies any Entropy *l*-diversity that one can impose), even though anyone in the equivalence class would be considered 98% positive, rather than 1% percent. In fact, this equivalence class has exactly the same diversity as a class that has 1 positive and 49 negative record, even though the two classes present very different levels of privacy risks.

5. (n, t)-Closeness: A NEW PRIVACY MEASURE

The (n, t)-closeness principle: An equivalence class E_1 is said to have (n, t)-closeness if there exists a set E_2 of records that is a natural superset of E_1 such that E_2 contains at least n records, and the distance between the two distributions of the sensitive attribute in E_1 and E_2 is no more than a threshold t. A table is said to have (n, t)-closeness if all equivalence classes have (n, t)-closeness.

The intuition is that to learn information about a population of a large-enough size (at least *n*). One key term in the above definition is "natural superset". Assume that we want to achieve (1000, 0.1)-closeness for the above example. The first equivalence class E_1 is defined by (zip code="476**", $20 \leq$ Age ≤ 29) and contains 600 tuples. One equivalence class that naturally contains it would be the one defined by (zip code= "476**", $20 \leq$ Age ≤ 39). Another such equivalence class would be the one defined by (zip code= "47***", $20 \leq$ Age ≤ 29). If both of the two large equivalence classes contain at least 1,000 records, and E_1's distribution is close to (i.e., the distance is at most 0.1) either of the two large equivalence classes, then E_1 satisfies (1,000, 0.1)-closeness.

In the above definition of the (*n*, *t*)-closeness principle, the parameter *n* defines the breadth of the observer's background knowledge. Smaller *n* means that the observer knows the sensitive information about a smaller group of records. The parameter *t* bounds the amount of sensitive information that the observer can get from the released table. A smaller *t* implies a stronger privacy requirement.

In fact, Table 5 satisfies (1,000, 0.1)-closeness. The second equivalence class satisfies (1,000, 0.1)-closeness because it contains 2,000 > 1,000 individuals, and thus, meets the privacy requirement (by setting the large group to be itself).

The first and the third equivalence classes also satisfy (1,000, 0.1)-closeness because both have the same distribution (the distribution is (0.5, 0.5)) as the large group which is the union of these two equivalence classes and the large group contains 1,000 individuals. Choosing the parameters *n* and *t* would affect the level of privacy and utility. The larger *n* is and the smaller *t* is, one achieves more privacy and less utility.

Table 4: Original Patients Table

	ZIP Code	Age	Disease	Count
1	47673	29	Cancer	100
2	47674	21	Flu	100
3	47605	25	Cancer	200
4	47602	23	Flu	200
5	47905	43	Cancer	100
6	47904	48	Flu	900
7	47906	47	Cancer	100
8	47907	41	Flu	900
9	47603	34	Cancer	100
10	47605	30	Flu	100
11	47602	36	Cancer	100
12	47607	32	Flu	100

Table 5: An Anonymous Version of Table 4

	ZIP Code	Age	Disease	Count
1	476**	2*	Cancer	300
2	476**	2*	Flu	300
3	479**	4*	Cancer	100
4	479**	4*	Flu	900
5	476**	3*	Cancer	100
6	476**	3*	Flu	100

6. ANONYMIZATION ALGORITHM

One challenge is designing algorithms for anonymizing the data to achieve (*n*, *t*)-closeness. This section, describes how to adapt the Mondrian multidimensional algorithm for (*n*, *t*)-closeness model. Since *t*-closeness is a special model of (*n*, *t*)-closeness, Mondrian can also be used to achieve *t*-closeness.

---

input: P is partitioned into r partitions $\{P_1, P_2, \ldots, P_r\}$

output: true if (*n*, *t*)-closeness is satisfied, false otherwise

---

for every P_i

if P_i contains less than *n* records find=false

for every Q ε Parent(P) and $|Q| \geq n$

if $D[P_i, Q] \leq t$, find=true

if find=false, return false

return true

Figure1. The Checking Algorithm

The algorithm consists of three components:
1) Choosing a dimension on which to partition;
2) Choosing a value to split; and
3) Checking if the partitioning violates the privacy requirement.

Figure 1 gives the algorithm for checking if a partitioning satisfies the (n, t)-closeness requirement. Let P be a set of tuples. Suppose that P is partitioned into r partitions {P1; P 2; . . . ; P r}, i.e., Ui {Pi}= P and Pi ∩ P j =Ø ; for any i ≠ j. Each partition Pi can be further partitioned and all partitions form a partition tree with P being the root. Let Parent (P) denote the set of partitions on the path from P to the root, which is the partition containing all tuples in the table. If Pi ($1 \leq i \leq r$) contains at least n records, then Pi satisfies the (n, t)-closeness requirement. If Pi ($1 \leq i \leq r$) contains less than n records, the algorithm computes the distance between Pi and each partition in Parent (P). If there exists at least one large partition (containing at least n records) in Parent (P) whose distance to Pi (D [Pi; Q]) is at most t, then Pi satisfies the (n, t)-closeness requirement. Otherwise, P_i violates the (*n*, *t*)-closeness requirement. The partitioning satisfies the (*n*, *t*)-closeness requirement if all P_i *s* have (*n*, *t*)-closeness

7. DISTANCE MEASUREMENT

Now, the problem is to measure the distance between two probabilistic distributions. There are a number of ways to define the distance between them. Given two distributions **P** = ($p_1, p_2, \ldots, p_m$), **Q** = ($q_1, q_2, \ldots, q_m$), two well-known distance measures are as follows:

The *variational distance* is defined as

$$D[P, Q] = \sum_{i=1}^{m} 1/2 \left| p_i - q_i \right|$$

And the Kullback-Leibler (KL) distance is defined as

$$D[P, Q] = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i} = H(P) - H(P, Q),$$

Where $H(P) = \sum_{i=1}^{m} p_i \log p_i$ is the entropy of **P** and $H(P, Q) = \sum_{i=1}^{m} p_i \log q_i$ is the cross entropy of **P** and **Q**.

These distance measures do not reflect the semantic distance among values. Recall Example 6 (Tables 2.6 and 2.7), where the overall distribution of the Income attribute is **Q** = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}. The first equivalence class in Table 2.7 has distribution $\mathbf{P_1}$ = {3k, 4k, 5k} and the second equivalence class has distribution $\mathbf{P_2}$ = {6k, 8k, 11k}. Our intuition is that $\mathbf{P_1}$ results in more information leakage than $\mathbf{P_2}$, because the values in $\mathbf{P_1}$ are all in the lower end; thus we would like to have $D[\mathbf{P_2}, \mathbf{Q}] > D[P_2; \mathbf{Q}]$. The distance measures mentioned above would not be able to do so, because from their point of view, values such as 3*k* and 6*k* are just different points and have no other semantic meaning.

In short, there is a metric space for the attribute values so that a ground distance is defined between any pair of values. Then there are two probability distributions over these values, and the distance between the two probability distributions to be dependent upon the ground distances among these values. This requirement leads to the Earth Mover's distance (EMD). EMD is described and how to use EMD in the closeness measures.

7.1 Earth Movers Distance

The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other. Intuitively, one distribution is seen as a mass of earth spread in the space and the other as a collection of holes in the same space. EMD measures the least amount of work needed to fill the holes with earth. A unit of work corresponds to moving a unit of earth by a unit of ground distance. EMD can be formally defined using the well-studied transportation problem. Let P = ($p_1, p_2, \ldots p_m$), Q = ($q_1, q_2, \ldots q_m$),and d_{ij} be the ground distance between element *i* of **P** and element *j* of **Q**. We want to find a flow $F = [f_{ij}]$, where f_{ij} is the flow of mass from element *i* of **P** to element *j* of **Q** that minimizes the overall work:

$$WORK(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij} ,$$

In the next section, formulas are derived for calculating EMD for the special cases that are needed to consider.

7.2 EMD for Numerical Attributes

Numerical attribute values are ordered. Let the attribute domain be $\{v_1, v_2 \ldots v_m\}$, where v_i is the ith smallest value. Ordered distance: The distance between two values of which is based on the number of values between them in the total order, i.e., ordered_ dist (v_i; v_j) = |i-j| /m-1 It is straightforward to verify that the ordered-distance measure is a metric. It is nonnegative and satisfies the symmetry property and the triangle inequality. To calculate EMD under ordered distance, it is only needed to consider flows that transport distribution mass between adjacent elements, because any transportation between two more distant elements can be equivalently decomposed into several transportations between adjacent elements. Based on this observation, minimal work can be achieved by satisfying all elements of **Q** sequentially. First consider element 1, which has an extra amount of $p_1 - q_1$. Assume, without loss of generality, that $p_1 - q_1 < 0$, an amount

of $q_1 - p_1$ should be transported from other elements to element 1. We can transport this from element 2. After this transportation, element 1 is satisfied and element 2 has an extra amount of $(p_1 - q_1) + (p_2 - q_2)$. Similarly, we can satisfy element 2 by transporting an amount of $|(p_1 - q_1) + (p_2 - q_2)|$ between element 2 and element 3. This process continues until element m is satisfied and **Q** is reached.

Formally, let $r_i = p_i - q_i$, (i = 1, 2. . . m), then the distance between P and Q can be calculated as

$$D[P, Q] = \frac{1}{m-1}\left(|r_1| + |r_1 + r_2| + \ldots + |r_1 + r_2 + \ldots r_{m-1}|\right)$$

$$= \frac{1}{m-1}\sum_{i=1}^{i=m}\left|\sum_{j=1}^{j=i} r_j\right|.$$

8. RESULTS

Here N can take any numeric value and T represents threshold value. For each group having less than N value is compared with all other groups and distance is calculated between them. If closeness is achieved for greater N value and small T value it represents more privacy.

AnonymousData

When the Anonymous Data button is clicked on the output screen it displays all the patients' records which are anonymized and grouped. After entering values for N and T click on NT Closeness button and save button, Then the EMD distance is calculated between the groups present in the anonymous table and the result is displayed as shown in the following figure

9. CONCLUSION

Initially the patients table having 60 records is created in the database. Then the quasi-identifier values in the table are replaced with the values that are less specific but semantically consistent. The table is anonymized which contains 9 groups i.e.., 27 records (9* 3=27). The proposed algorithm is implemented on the anonymized table to calculate the distance between the groups. A series of values for *n* ant *t* are taken to show the capability of the proposed algorithm

10. SCOPE FOR FUTURE WORK

Multiple sensitive attributes present additional challenges. Suppose if there are two sensitive attributes U and V. One can consider the two attributes separately, i.e., an equivalence class E has (*n*, *t*)-closeness if E has (*n*, *t*)-closeness with respect to both U and V. Another approach is to consider the joint distribution of the two attributes. To use this approach, one has to choose the ground distance between pairs of sensitive attribute values. A simple formula for calculating EMD may be difficult to derive, and the relationship between (*n*, *t*) and the level of privacy become more complicated.

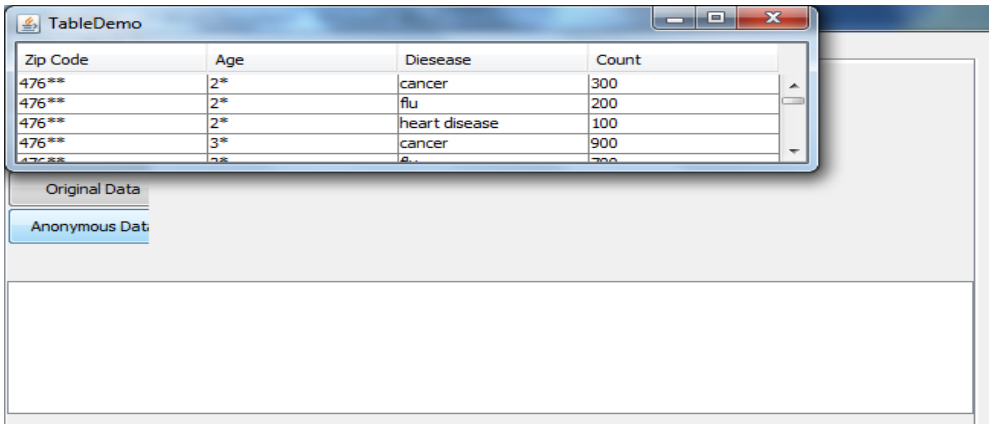

Fig 1. Anonymous Data Table

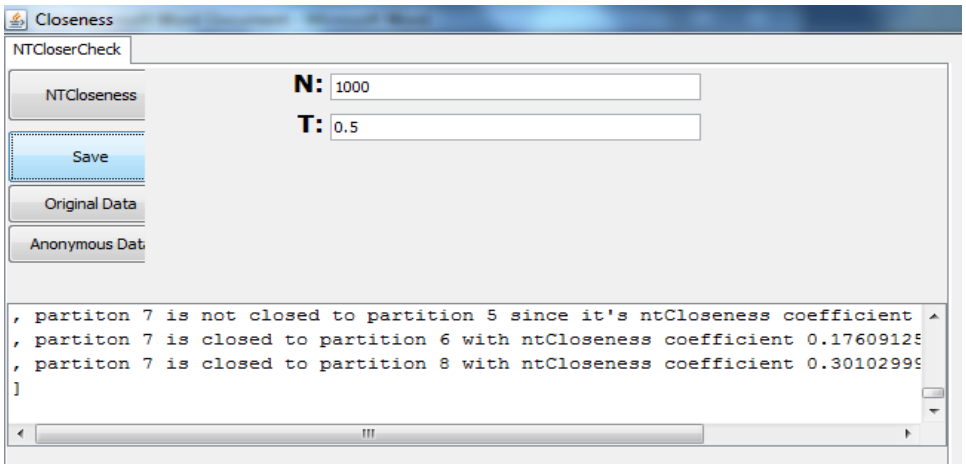

Fig.2. The closeness can be verified by giving different values for N and T

11. REFERENCES

[1] N. Li, T. Li, and S. Venkatasubramanian, "Closeness: A New Privacy Measure for Data Publishing," Proc. Int'l Conf. Data Eng. (ICDE), pp. 943-956, 2010.

[2] D. Lambert, "Measures of Disclosure Risk and Harm," J.OfficialStatistics, vol. 9, pp. 313-331, 1993.

[3] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l pp. 557-570, J.Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, 2002.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M.Venkatasubramaniam, "'l-diversity: Privacy Beyond k- Anonymity," Proc. Int'l Conf. Data Eng. (ICDE), pp. 24, 2006

[5] N. Li, T. Li, and S. Venkatasubramanian, "*t*-Closeness: Privacy beyond *k*-Anonymity and *l*- Diversity," Proc. *Int'l Conf. Data Eng. (ICDE)*, pp. 106-115, 2007.

[6] K. LeFevre, D. DeWitt, and R. Ramakrishnan "Mondrian Multidimensional-Anonymity," Proc. Int'l Conf. Data Eng. (ICDE),p. 25, 2006.

[7] N. Li, T. Li, and S. Venkatasubramanian, "t- Closeness: Privacy beyond k-Anonymity and 'Diversity," Proc. Int'l Conf. Data Eng. (ICDE), pp. 106-115, 2007.

[8] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,"'-Diversity: Privacy Beyond k-Anonymity," Proc. Int'l Conf.Data Eng. (ICDE), p. 24, 2006.

[9] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'lJ. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[10] T. Li and N. Li, "Towards Optimal k- Anonymization," Data and Knowledge Eng., vol. 65, pp. 22-39, 2008.

[11] R.J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization," Proc. Int'l Conf. Data Eng. (ICDE), pp. 217-228, 2005.

[12] B.C.M. Fung, K. Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation," Proc. Int'l Conf. Data Eng. (ICDE), pp. 205-216, 2005.

[13] T. Li, N. Li, and J. Zhang, "Modeling and Integrating Background Knowledge in Data Anonymization," Proc. Int'l Conf. Data Eng. (ICDE), 2009.

[14] M.E. Nergiz, M. Atzori, and C. Clifton, "Hiding the Presence of Individuals from Shared Databases," Proc. ACM SIGMOD,pp. 665-676, 2007.

[15] H. Park and K. Shim, "Approximate Algorithms for K-Anonymity,"Proc. ACM SIGMOD, pp. 67-78, 2007.

[16] V. Rastogi, S. Hong, and D. Susie, "The Boundary between Privacy and Utility in Data Publishing," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 531-542, 2007.

[17] R.C.-W. Wong, J. Li, A.W.-C. Fu, and K. Wang, "k-Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing," Proc. ACM SIGKDD, pp. 754-759, 2006.

[18] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. Int'l Conf. Very Large Data Bases (VLDB),pp. 139-150, 2006.

[19] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc.ACM SIGMOD, pp. 229-240, 2006.

[20] X. Xiao and Y. Tao, "m-Invariance: Towards Privacy Preserving Republication of Dynamic Datasets," Proc. ACM SIGMOD,pp. 689-700, 2007.

[21] V.S. Iyengar, "Transforming Data to Satisfy Privacy Constraints,"Proc. ACM SIGKDD, pp. 279-288, 2002.

[22] D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Datasets," Proc. ACM SIGMOD, pp. 217-228, 2006.

[23] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, "Aggregate Query Answering on Anonymized Tables," Proc. Int'l Conf. Data Eng.(ICDE), pp. 116-125, 2007.

AUTHOR'S PROFILE

A.Sunitha obtained her Bachelor Degree in Computer Science and Engineering from JNTU Hyderabad, A.P., India in 2000, and obtained M.Tech Computer Science in 2011 from JNTUH College of Engineering, Hyderabad. Her Research interest includes Data Mining, Database Systems and Computer Networks. Currently she is working as Assistant Professor in CSE Department at St.Martin's Engineering College, Hyderabad.

K.Venkata Subba Reddy obtained his B.Tech in Information Technology from University of Madras in 2002 and received the M.Tech in Software Engineering from Bharath University, Chennai in 2005, He is currently pursuing Ph.D., in Computer Science and Engineering, under the guidance of Dr.B.Raveendra babu, at Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. His research interests include Software Engineering, Database Systems and Cloud computing. He is a life member of ISTE and a member of CSI. Currently he is working as Assistant Professor in Computer Science & Engineering Department at Muffakham Jah College of Engineering & Technology, Banjarahills, Hyderabad.

B Vijaya Kumar Completed his M S in CSE from DPI, Donetsk, USSR in 1993. He is pursuing Ph.D. in Computer Science & Engineering under the guidance of Dr.Vakulabharanam Vijayakumar. He is a life member of CSI, ISTE, NESA and ISCA.He is working as Professor & HOD of IT & MCA Dept at St.Martin's Engineering College, Secunderabad, India. He has published more than 10 research publications in various National, International Journals.