L Diversity on K-Anonymity with External Database for improving Privacy Preserving Data Publishing

P. Mayil Vel Kumar
Research scholar
Anna University
Chennai

M. Karthikeyan, PhD.
Professor & HOD (ECE)
Tamilnadu College of Engg.
Coimbatore

ABSTRACT
The data must be secure and measurable at the public when it releases to view. The data table produces personal information and sensitive values. They are maintained for secrecy, the anonymity is the best method to protect the data. There are many anonymity methods to protect the data. k -anonymity is one method to protect the data. The problem in k- anonymity method is if data set increases then utility decreases. Also k-anonymity data is possible to many attacks like Homogeneity Attack, Background Knowledge Attack. The ℓ -diversity is another method to protect the data. Main advantage of ℓ- diversity is the data set increases then the data utility also increases. Based on above advantage, we applied ℓ -diversity concept in k-anonymity applied external data set and we evaluate high efficiency dataset. It shows the ℓ -diversity reduces the data losses in k anonymity data sets when data point moves any size.

General Terms
Data mining, privacy, anonymity, Security, Algorithms.

Keywords
Data pre-processing, k-anonymity, **ℓ** –diversity, quasi-identifier.

1. INTRODUCTION
Various organizations (e.g., Hospital authorities, industries and government organizations etc) releasing person specific data, which called as micro data. They provide information of privacy of individuals. The main aim of privacy is to protect information, at the same time the data must produce external knowledge. Based on data value, it is divided into three types 1) identifiers 2) quasi-identifiers and 3) sensitive attributes. This micro data's consists in the form of a table which is called as micro table. In this micro table the identifiers (e.g., employ id and names) can be used individually to identify a table, so they must completely remove. Quasi-identifiers (e.g., date of birth and gender) partially hidden data. Sensitive attributes (e.g., disease, salary, and criminal offence) field should not hide, since they can produce external knowledge. The process of concealing identity information in micro data is called de identification. On the other hand, re-identification is the successful linking of a published table to an existing person and corresponds to a privacy breach. Privacy Preserving Data Mining performs data mining on the private data. Different methods such as anonymization, perturbation and cryptographical approaches have been used for privatizing the data. Here anonymization methods are taken to protect the data. Very early stage the k-anonymity privacy methods are used for publishing micro data [1] [3] [4] [8]. Recently, several authors have recognized that k-anonymity cannot prevent attribute disclosure [2] [5] [6] [7] [10]. So the notion of ℓ-diversity has been proposed to address this data [2]. Here a complete ℓ - diversity applied k- anonymity external data set model is proposed which can implement sensitive values individuation preservation by setting the frequency constraints for each sensitive value in all the equivalence class.

Table 1. Patients List

	Non-Sensitive			Sensitive
	Zip Code	**Age**	**Nationality**	**Condition**
1	13053	28	Russian	HIV
2	13068	29	American	HIV
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	HIV
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

Sweeney [3] proposed the k-anonymity model, where some of the quasi-identifier fields are suppressed or generalized[4] 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. The Table 1 shows patient lists and Table 2 shows a 2-anonymous view corresponding to it. The sensitive attribute (Health Condition) is retained without change in this example. In the literature of k-anonymity problem, there are two main models. One model is called global recoding and second is local recoding. Here, we assume as each attributes are a corresponding conceptual generalization hierarchy or taxonomy tree. As per hierarchy the lower level domain has more details than higher level domain. For example, Zip Code 13053 is a lower level domain and Zip Code 130** is a higher level domain. We assume such hierarchies for numerical attributes too. In particular, we have a hierarchical structure defined with {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, Age 28, 29 in the lower level can be replaced by the interval (<40) in the higher level (See Table 2).In recent years numerous algorithms [1][4] have been proposed for implementing k-anonymity via generalization and

suppression. Each generalized and suppressed data's are grouped constrains as on same size in each cluster.

2. OUR CONTRIBUTIONS

In this paper, we focus on the using generalization and suppression [4] method. While focusing on identity disclosure, k-anonymity model fails to protect attribute disclosure [1]. Several models such as p-sensitive, k-anonymity [3], ℓ -diversity [2], and t-closeness [5] [9] 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 [1].The main contribution of [1] is to introduce to apply the ℓ -diversity [2] in k-anonymity property data set, which requires, in addition to k-anonymity, that for each group of table with identical combination of quasi-identifier values, the number of distinct sensitive attributes values must be at least l. However, depending on the nature of the sensitive attributes even ℓ -diversity property still permits the information to be disclosed.

The proposed new privacy protection models called ℓ-diversity on k- anonymity applied on micro data with external database to reduce information loss of dataset while maintaining the same security of disclosure attack. It is found that if the number of quasi-identifiers increases, the balancing point moves down in k anonymity applied data base. So, k-privacy datasets are increased but utility decreased. The ℓ -data provides high efficiency of data sets when data point moves any size. Previously authors used separately k-anonymity and ℓ- diversity principles, we applied this two concept in data sets. The experiment conducted and the result for ℓ- diversity on k- anonymity applied external data model implemented.

Table2. 2 Anonymity Table

	Non-Sensitive			Sensitive
	Zip Code	**Age**	**Nationality**	**Condition**
1	130**	<30	*	HIV
2	130**	<30	*	HIV
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	1485*	>40	*	Cancer
6	1485*	>40	*	HIV
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

3. PROBLEM DEFINITION

3.1 Definition 1 (Micro data)

Attribute which must not be disclosed in the released micro data.

3.2 Definition 2 (Sensitive attribute)

The data to be released after applying anonymization methods on it is called the sensitive attribute.

3.3 Definition 3 (Quasi-Identifier)

The table T with attributes ($A_1 \ldots A_n$), a quasi identifier is a minimal set of attributes ($A_{i1} \ldots A_{il}$) ($1 <= i_1 < \ldots < i_l <= n$) in T that can be joined with external information to re-identify individual records.

3.4 Definition 4(Equivalence Class)

All set of tables which cannot be distinguished from each other with respect to Quasi-Identifier are called an Equivalence class

3.5 Definition 5 (K-Anonymity)

A table T is said to be k anonymous given a parameter k and the quasi-identifier QI = ($A_{i1} \ldots A_{il}$) if for each table $t_i \in T$, there exist at least another (k-1) table's $t_1 \ldots t_{k-1}$ such that those k tables have the same projection on the quasi-identifier. Table t and all other tables indistinguishable from t on the quasi-identifier form an equivalence class.

Based on above definition K-anonymity works with two concepts Generalization and Suppression [4]. Suppression is masking the attribute value with a special value in the domain. Generalization is replacing a specific value with a more generalized one. K-anonymity for every single attribute in QI can combine with each other by generalization and suppression, until all set of attributes to make considerable value. As with these values the k-anonymity data in table with a specific value for the quasi-identifier have the same sensitive attribute value. These values can easily attacked by external source.

4. ATTACKS ON K ANONYMITY

In this section we present two attacks, the homogeneity attack and the background knowledge attack, and we show how they can be used to compromise a k-anonymous dataset.

4.1 Homogeneity Attack

Alice and Bob are antagonistic neighbours. 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 cancer. Alice discovers the 2-anonymous table of current inpatient records published by the hospital (table 2), 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 5, 9, 10, 11, or 12. Now, all of those patients have the same medical condition (cancer), and so Alice concludes that Bob has cancer.

4.2 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 tables where the sensitive attribute can take 3 distinct values and is not correlated with the any 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 tables that share the same values of their quasi-identifiers should have diverse values for their sensitive

attributes. Our next observation is that an adversary could use "background" knowledge to discover sensitive information.

4.3 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 table 2. 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 HIV. 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.

4.4 Observation 2

k -Anonymity does not protect against attacks based on background knowledge. It demonstrated (using the homogeneity and background knowledge attacks) that a k-anonymous table may disclose sensitive information. Since both of these attacks are plausible in real life, we need a stronger definition of privacy that takes into account diversity and background knowledge. This paper addresses this very issue.

To avoid the above problems of k-anonymity attacks the ℓ-diversity concept was introduced

4.5 L-Diversity

The drawback of k-anonymization due to the background knowledge attack can be removed by diversifying the values of sensitive attribute within a block. The ℓ -diversity model is a very useful model for preventing attribute disclosure and it has been introduced in [2].

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.

The entropy of an equivalence class E is defined to be Entropy(E) = $-\Sigma_{s\epsilon S}$ 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. As with above definition the sensitive attributes QI can well diverse by each group, so it does not take into account the semantically closeness of these values.

4.6 Properties

- Knowledge of the full distribution of the sensitive and non-sensitive attributes is not required in ℓ -diversity.
- ℓ -diversity does not even require the data publisher to have as much information as the adversary. The larger the value of l, the more information is needed to rule out possible values of the sensitive attribute.
- Different adversaries can have different background knowledge leading to different inferences. It simultaneously protects against all of them without the need for checking which inferences can be made with which levels of background knowledge.

4.7 Distinct L-diversity

The term "well represented" in the definition of ℓ -diversity would be to ensure there are at least ℓ distinct values for the sensitive attribute in each equivalence class. Distinct ℓ -diversity does not prevent probabilistic inference attacks. It may happen that in an anonymized block 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.

4.8 Entropy L-diversity

The entropy of an equivalence class E is defined to be

$$E = -p(E, s) \log p(E, s)$$

Where 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 strong than distinct ℓ -diversity. 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.

4.9 Recursive (c, ℓ)-diversity

Recursive (c, ℓ)-diversity ensure 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 ith most frequent sensitive value appears in an equivalence class E. Then E is said to have recursive (c, ℓ)-diversity if $r_1 < c(r_l + r_{l+1} + ... + r_m)$. A table is said to have recursive (c, ℓ)-diversity if all of its equivalence classes have recursive (c, ℓ)-diversity.

5. EXAMPLE OF L DIVERSITY ON k ANONYMITY DATASET

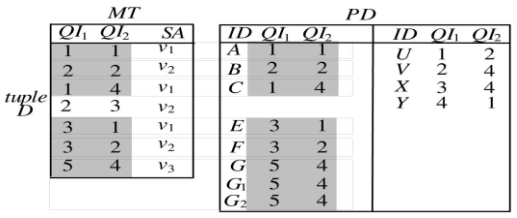

Fig 1a&b: micro table and public data

The figure 1a shows well represented medical record as called public data with two quasi identifier and they grouped as A, B, C, E, F and G. Moreover, it includes six additional records: G1, G2 (which have identical QI values to G), U, V, X, and Y. The generalization applied to the public data this convert as micro table (MT) consist of group various groups. MT in Fig. 1b is 1-anonymous as all combinations of QI values are distinct. The process of generating a k-anonymous table given the original micro data is called k-anonymization. The most common form of k-anonymization is generalization, which involves replacing specific QI values with more general ones.

5.1 Construction of G box

The most common method, i.e., mapping, for achieving anonymization is generalization. For numerical QIs, a generalization of a value is a range. For categorical QIs, it is a higher level value in a given hierarchy (e.g., a city name is replaced with a state or country). Since categorical values can be trivially mapped to an integer domain, we assume only numerical QIs here. A generalized table is represented as an axis-parallel (hyper) rectangle, called G-box, in the QI space defined by the extent of its QI ranges. We use the term anonymized group, or simply group, to refer to the set of MT table that fall within a G-box. The goal of k-anonymity is to hide the identity of individuals by constructing G-boxes that contain at least k MT table.

5.2 Implementation of Mondrian

In this module, Mondrian constructs QI groups than contain from k up to 2k - 1 table (when all QI values present in MT are distinct), following a strategy similar to the KD-tree space partitioning. In particular, starting with all MT records, it splits the d-dimensional space (defined by the d QI attributes) into two partitions of equal cardinality. The first split is performed along the first dimension (i.e., quasi-identifier QI_1, according to the median QI_1 value in MT. Each of the resulting groups is further divided into two halves according to the second dimension. Partitioning proceeds recursively, choosing the splitting dimension in a round-robin fashion among QI attributes. Mondrian terminates when each group contains fewer than 2k records. The resulting space partition is the anonymous version of MT to be published.

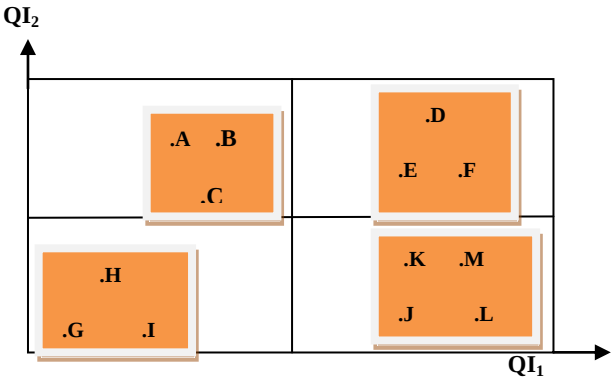

Fig2. Generalization of MT with Mondrian.

Fig. 2 demonstrates 3-anonymization with Mondrian, assuming that MT contains records A . . . M and has two quasi-identifiers. The horizontal axis corresponds to QI_1, and the vertical to QI_2. The first split is performed on the horizontal axis, according to the QI_1 value of C. The left (right) half of the space contains 6 (7) MT table (i.e., exceeding 2k - 1 = 5), and it is divided into two groups according to the QI_2 value of record A (of record F). Since each resulting group has fewer than 5 tables, splitting terminates. The anonymized version AT of MT consists of the four shaded minimum bounding boxes (MBBs), each representing an anonymized group.

5.3 Implementation of top down

In this module, Top Down is a recursive clustering algorithm. Specifically, it starts with the entire MT and progressively builds tighter clusters with fewer points. Fig. 3 demonstrates the steps of Top Down on the MT table of Fig. 1. Initially, the algorithm finds the 2 table that if included in the same anonymized group, they would result in the largest perimeter. In our example, this first step retrieves G and E. Next, Top Down considers the remaining records in random order, and groups them together with either G or E; a considered table is inserted to the group where it causes the smallest NCP increase.

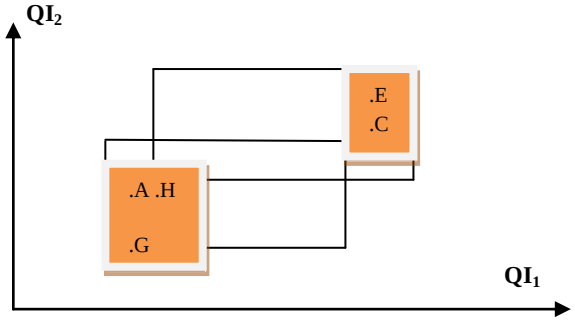

Fig. 3. Generalization of MT with TopDown.

In Fig. 3, assume that record A is processed first. It is included in G's cluster, because if grouped with E, it would lead to a rectangle with larger perimeter. Similarly, if C (H) is the second tuple, it is grouped with E (with G and A). After the first pass, all records belong to either group. The procedure is repeated recursively within each cluster, until all groups have no more than k tuples. After this step, the majority of the groups have cardinality below k. To fulfill the k-anonymity requirement, undersized groups are merged with neighboring ones according to some heuristics, aiming at a small NCP. The shaded MBBs in Figure (4) below correspond to four anonymized groups in our example.

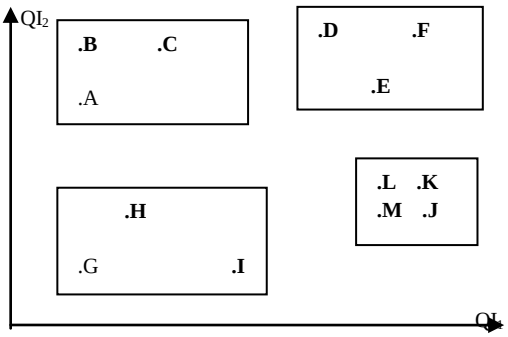

Fig. 4 Grouped of MT with Top Down

5.4 Apply L-diversity on k-anonymity data set

To improve performance of k-anonymity and to avoid the various attacks, the ℓ -diversity applied principal in shown below.

In our example, consider the inpatient records shown in table 1 have changed has a 3-diverse version of the table in table 3. Comparing it with the 2-anonymous table 2 we see that the attacks against the 2-anonymous table are prevented by the 3-diverse table. For example, Alice cannot infer from the 3-diverse table that Bob (a 31 year old American from zip code 13053) has cancer. Even though Umeko (21 year old Japanese from zip code 13068) is extremely unlikely to have HIV, Alice is still unsure whether Umeko has a viral infection or cancer. The ℓ-diversity principle advocates ensuring ℓ "well represented" values for the sensitive attribute in every q*-

block, but does not clearly state what "well represented" means. Note that we called it a "principle" instead of a theorem we will use it to give two concrete instantiations of the ℓ-diversity principle and discuss their relative trade-offs. To make the instantiation of ℓ-diversity principle theoretic notion of entropy taken for every q^*-block

$$-\sum_{s \in S} P(q * s') \log(P(q * s')) \geq \log(l)$$

Based on this definition every q* block contains ℓ distinct values for every sensitivity attribute Using this definition, table 3 is actually 1.4 –diverse.

Since x log(x) is a concave function, it can be shown that if we split a q*block into two sub-blocks q_a*and q_b* then entropy (q*) ≥ min (entropy (q_a*), entropy (q_b*). This implies that in order for entropy ℓ-diversity to be possible, the entropy of the entire table must be at least log (ℓ). This might not be the case, especially if one value of the sensitive attribute is very common, for example, if 90% of the patients have "cancer" as the value for the "Medical Condition" attribute.

Table3. 3-Diverse Inpatient Micro data

	Non-Sensitive			Sensitive
	Zip Code	**Age**	**Nationality**	**Condition**
1	1305*	≤ 40	*	HIV
4	1305*	≤ 40	*	Viral Infection
10	1305*	≤ 40	*	Cancer
9	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
7	1485*	> 40	*	Viral Infection
6	1485*	> 40	*	HIV
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	HIV
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Thus entropy ℓ-diversity may sometimes be too restrictive. If some positive disclosures are acceptable (for example, a clinic is allowed to disclose that a patient has a "Viral Infection" because it is well known that most patients who visit the clinic have viral problems) then we can do better. This reasoning allows us to develop a less conservative instantiation of the ℓ-diversity principle called recursive ℓ-diversity.

Let S_1,…..,S_m be the possible values of the sensitive attribute S in a q* -block. Assume that we sort the counts n (q*, s_1) . . . , n (q*, s_m) in descending order and name the elements of the resulting sequence r_1, . . . , r_m. One way to think about ℓ-diversity is that the adversary needs to eliminate at least ℓ − 1 possible values of S in order to infer a positive disclosure. This means that, for example, in a 2-diverse table, none of the sensitive values should appear too frequently. We say that a q* -block is (c, 2)-diverse if $r_1 < c (r_2 + \ldots\ldots + r_m)$ for some user-specified constant c. For ℓ > 2, we say that a q*-block satisfies recursive (c, ℓ) - diversity if we can eliminate one possible sensitive value in the q* -block and still have a (c, ℓ−1)-diverse block.

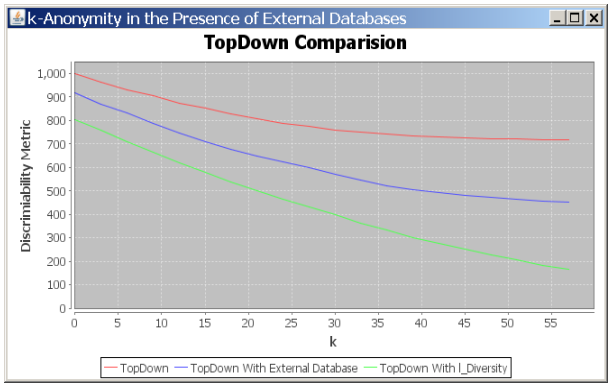

Fig 5: public data in top down comparison curve graph

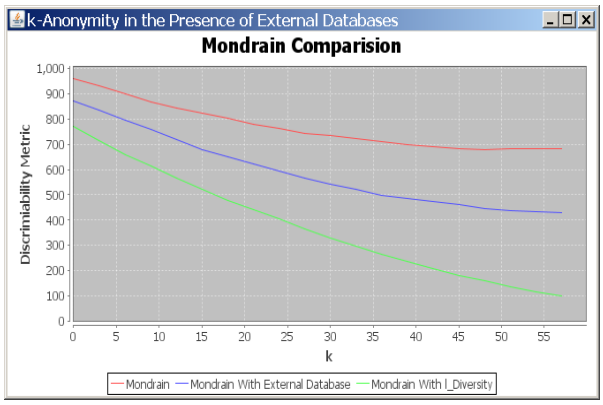

Fig 6: public data Mondrian comparison curve graph

6. EXPERIMENTS

In our experiments, we used an implementation of Incognito, as described in [1], for generating k-anonymous tables than we modified this implementation with ℓ-diverse concept as in [2]. So that it produces ℓ-diverse tables as well defined data set the performance shown in figure (5, 6, 7&8). Incognito is implemented in Java and uses the database manager IBM DB2 to store its data. All experiments were run under Windows XP on a machine with a 3 GHz Intel Pentium 4 processor and 1 GB RAM. We ran our experiments on the Adult Database from the UCI Machine Learning Repository and the Lands End Database. The Adult Database contains 10000 tables from US Census data and the Lands End Database contains 5,000,000 table of point-of-sale information. We removed table with missing values and adopted the same domain generalizations as [1]. We used the number of distinct values for each attribute, the type of generalization that was used (for non-sensitive attributes), and the height of the generalization hierarchy for each attribute. Due to space restrictions, we report only a small subset of our

experiments.

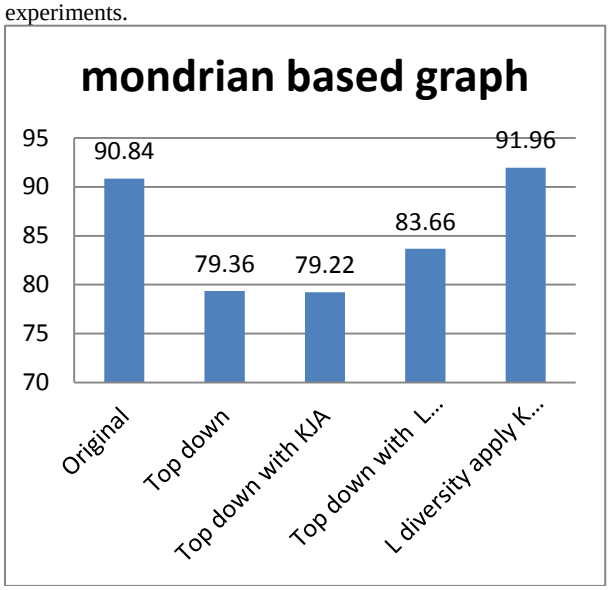

Fig 7: public data in Mondrian comparison step graph

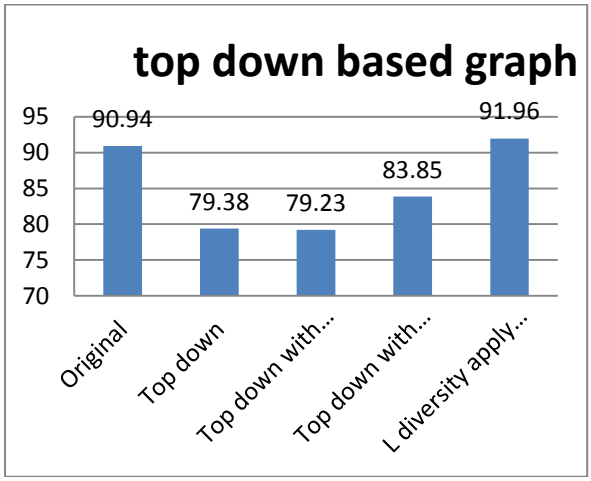

Fig 8: public data in top down comparison step graph

7. PERFORMANCE

In our set of experiments, we compare the running times of entropy ℓ-diversity and k-anonymity. The results are shown in Figures 5, 6,7and 8. For the Adult Database, we used Occupation as the sensitive attribute, and for Lands End we used Cost. We varied the quasi-identifier size from 3 attributes up to 8 attributes; a quasi-identifier of size j consisted of the first j attributes of its dataset as listed in Figure 5, 6, 7&8. We measured the time taken to return all 3-anonymous tables and compared it to the time taken to return all 3-diverse tables. In both datasets, the running times for k-anonymity and ℓ-diversity were similar. Sometimes the running time for ℓ-diversity was faster, which happened when the algorithm pruned parts of the generalization lattice earlier than it did for k-anonymity.

8. RESULT

For evaluating utility, we performed the classification mining on the ℓ-diversity and k-anonymized dataset. Classification was performed by using jdk7.0 Software considering native-country as classification variable. We considered the percentage of correctly classified table as the utility of the dataset. Figure 5, 6, 7 and 8 shows the results produced by the jdk7.0 on using ℓ-diversity on k-anonymity for an anonymized dataset in Mondrian and Top down approach. Privacy was calculated by counting the number of table which are generalized with anonymized values. Privacy and utility were calculated by varying the value of ℓ and k. The balancing point between utility and privacy is the point where privacy and utility curves intersect or tend to converge. Figure 5&6 shows the variation of utility and privacy with k, ℓ value applied for Mondrian & top down approach. It clearly follows from the figure (5, 6, 7&8) that on increasing the value of k privacy provided by the dataset increases but utility decreases. For this sample dataset the balancing point comes between k=5 and ℓ =3, and utility of the dataset at balancing point is around 90%.

Table4. Comparison of privacy models

Privacy Model	Run Time	Balance Point	Data Utility	Data Accuracy
k - anonymity	Low	Increase	Decrease	Medium
ℓ - diversity	High	Increase	Increase	Compare to Previous, This is High
ℓ - diversity applied k - anonymity external data model	Very High	Increase	Increase	Compare to Previous two, This is High

In order to improve the privacy offered by the dataset, utility of the data suffers. On conducting the experiments we found that the balancing point between utility and privacy depends on the dataset and value of k cannot be generalized for all datasets such that utility and privacy are balanced. On varying the number of sensitive attributes in a dataset the balancing point varies. We found that if the number of quasi-identifiers increases, the balancing point moves down and balance between utility and privacy occurs at a higher value of k. Thus if a dataset contains more number of quasi identifiers then the utility as well as privacy attained at balancing point will be less than the dataset having fewer quasi identifiers. We also studied the effect of number of table in the data set on the balancing point and found that as the number of table increases there is slight shift in the balancing point and the value of k for which balancing occurs. Thus we can approximately predict the balancing point for a huge dataset by conducting experiment on a sample dataset. It shows k anonymity value is 90.84% but same data set with apply of ℓ diversity is 91.96% for Mondrian approach. In our second approach shows k anonymity value is 90.94% but same data set with apply of l diversity is 91.96% for top down approach. So efficiency of data set is increased. Our result performance comparison is shown in table 4.

9. CONCLUSION AND FUTURE WORK

In this paper, we introduced a new approach for privacy preserving for ℓ-diversity on k-anonymity with knowledge of external database. Previously the authors measure the data sets separately as in [1] & [2]. The data set taken is in generalization boundaries imposed by the data which in quasi identifier. An experimental evaluation indicates that often the result of ℓ diversity on k- anonymity data set has better solution than k-anonymity data set. The ℓ -diversity concept is alone not efficient for protecting data in more dense datasets. We can further investigate an approach for applying techniques on k-anonymity to protect more dense datasets.

10. REFERENCES

[1] Dimitris Sacharidis, Kyriakos Mouratidis, and Dimitris Papadias.K-Anonymity in the presence of External database, IEEE Transactions on Knowledge and Data Engineering, vol.22, No.3, March 2010

[2] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, Muthuramakrishnan Venkitasubramaniam, ℓ-Diversity: Privacy Beyond k-Anonymity

[3] Sweeney.L, k-anonymity: a model for protecting privacy. International Journal on Uncertainty,Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[4] Sweeney, L. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression.International Journal of Uncertainty, Fuzziness and Knowledge-Based System, 10(5) pp. 571-588, 2002.

[5] Li, N. Li, T. Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. ICDE 2007: 106-115

[6] Poovammal.E,Ponnavaikko.M, Privacy and Utility preserving Task Independent Data Mining International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 15

[7] Sowmyarani C N, G N Srinivasan Survey on Recent Developments in Privacy Preserving Models International Journal of Computer Applications (0975 – 8887) Volume 38– No.9, January 2012

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

[9] Li, N. Li, T. Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. ICDE 2007: 106-115

[10] Sun, X. Wang, H. and Li, J. On the complexity of restricted k-anonymity problem. The 10th Asia Pacific Web Conference (APWEB2008), LNCS 4976, pp: 287-296, Shenyang, China.