Privacy Issues for K-anonymity Model

Nidhi Maheshwarkar*, Kshitij Pathak, Vivekananad Chourey*****

*, ** (Department of Information Technology,
Mahakal Institute of Technology, Ujjain, India)
*** (Department of Digital Communication
Mandsaur Institute of Technology,
Mandsaur, India)

Abstract—

K-anonymity is the approach used for preventing identity disclosure. Identity disclosure means an individual is linked to a particular record in the published data and individual's sensitive data is accessed .Some important information such as Name, Income details , Medical Status and Property details are considered as a sensitive data(or Attribute) because these data have to be kept secure from unauthorized access. Generally these details are stored in private tables of any organization or committees. Some released attributes called as quasi identifiers (Zip code, Sex, marital status, Age, Date of Birth, Bank details) when linked with private table cause the Identity disclosure. In this paper we will discuss some privacy issues for k-anonymity model and check its integrity while using some approaches.

Keywords- K-anonymity model ,Attacks ,l-diversity,t-closeness, Sensitive tuples.

I. INTRODUCTION

Data privacy is the major issue of today's global world, where internet has many advantages in the sector of education, communication, online medical help etc. But it has some disadvantages regarding data privacy. Much sensitive information can be traced via internet. Huge amount of data regarding any Organization or Committee, generally stored in the form of table and this table considered as a private table of Organization. The people belongs from this organization obviously have the part of that country or city. So, some information regarding them is available in released data which are publically available. It's not a tuff task for an adversary to access individual's information. If we want to search any details regarding any person (as we don't have any evil intension), then if we type his name in Google (any search engine), we get some information about him which is publically available. Then we can think an adversary who access data using almost every possible links can use this data for his benefit.

As shown in Fig. 1 PT is a public table whose attributes are { Name, Zip code, Age, Marital status, Nationality } have values for a tuples {Jack, 14853, 50, Indian}when linked with Fig 2. CT attributes {Zip code, Date of Birth, Race} who have same set of values disclosed that Jack has Cancer. This type of Attack is known as linking attack. Therefore we need a new privacy model which prevent from this linking attack. Sweeny [2002] gives a new privacy model known as K-anonymity model which prevent this type of attacks, and when we add some new concept like *l*-diversity, t-closeness, its privacy criteria increases and data will be protected by other major attacks like Homogeneity attack and Background knowledge attack. We will discuss K-anonymity model in II section and attack, *l*-diversity and t-closeness in respectively III and IV section.

NAME	ZIPCODE	AGE	MARITAL STATUS	NATIONALITY
....	....	....	....	....
....	....	....	....	....
....	....	....	....	....
....	....	....	....	....
Jack	14853	50	Single	Indian
....	....	....	....	....
	....	....	....	....
....	....	....	....	....
....	....	....	....	....

Fig 1 Public Table[PT]

S. NO	NONSENSITIVE			SENSITIVE
	ZIP CODE	AGE	NATIONALITY	MEDICAL STATUS
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

Fig. 2 Inpatient Microdata [CT]

II. K-ANONYMITY MODEL

Anonymity refers to a nameless state, a state where a data doesn't show its identity's. K-anonymity emphasizes that each released record has at least (K-1) other records in the release whose values are indistinct over those fields that appear in external 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 the Linking attacks [5].

For example if an adversary David knows {Zip code, Age, Nationality} = {13053, 28, Russian} of Tom. When he links these details with private table as shown in Fig 3 he found that there are 4 people present who show these details so he was fail to detect tom Medical status. So, K-anonymity diverts the concentration of adversaries and prevents identity disclosure. To make a k-anonymous table one assumption is needed, the data holder knows which attributes may appear in the external information and possibly is available to the data recipients and, therefore, which sets of attributes are quasi-identifiers.

To achieve K-anonymity two popular approaches used named as Generalization and Suppression .In generalization we replace a value with '*' or any less specific and more general value. For example Zip code 10353 will replace with 1035* so it will show any value from 10350 to 10359 if we replace second last digit with * then we get 103** so it will work for 10300 to 10399 and an adversary will be confused and cant infer actual data. In the case of age we can replace it with * and also show using range or other forms. For example 28 can be replaced by 2* or <30, [20-29].

Next approach is Suppression The intuition behind the introduction of suppression is that this additional method can reduce the amount of generalization necessary to satisfy the k-anonymity constraint. Suppression is therefore used to "moderate" the generalization process when a limited number of outliers (i.e., tuples with less than k occurrences) would force a great amount of generalization [1].In fig 3 * in Nationality shows tuples suppression. Others Technique used to achieve anonymity sampling, swapping values, randomization, and adding noise to data.

III ATTACKS AND PREVENTION

In this section we present two major attacks, the homogeneity attack and background knowledge attack, along with unsorted matching attack, complementary release attack and temporal attack, and we show that how they can be used to compromise a k-anonymous dataset. So here new definition arise *l*-diversity. ℓ-Diversity provides privacy even when the data publisher does not know what kind of knowledge is possessed by the adversary. The main idea behind ℓ-diversity is the requirement that the values of the sensitive attributes are well-represented in each group.

S. NO	NONSENSITIVE			SENSITIVE
	ZIP CODE	AGE	NATIONALITY	MEDICAL STATUS
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

Fig. 3 4-Anonymous Inpatient Microdata

Even when sufficient care is taken to identify the QI, the k-anonymity is still vulnerable to attacks. The common attacks are unsorted matching attacks, complementary release attacks and temporal attacks. Fortunately, these attacks can be prevented by some best practices. But the two major attacks, Homogeneity and Background attacks disclose the individuals' sensitive information. K-anonymity does not protect against attacks based on background knowledge because k-anonymity can create groups that leak information.

Observation 1: K-anonymity does not provide privacy in case of Homogeneity and Background attacks.

Homogeneity Attack: Suppose A and B are enemies and A wants to infer B's medical status which is present in fig. 4. A knows B's ZIP code is 13053 and his age is 35. So using this knowledge A knows that B's records belong from record no. 9,10,11,12 have Cancer. So A concludes that B has Cancer. This situation or attack is implies that k-anonymity can create groups which are responsible for leakage of information. This happens due to the lack of diversity in the sensitive attribute. This problem suggests that in addition to k-anonymity, the disinfected 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: Suppose C and D are two aggressive neighbors and C wants to infer D's private data, let the medical status, from the private table PT. Figure 4 shows a 4-anonymous private table with patient micro data which satisfies k-anonymity. So for a single value, C finds 3 more values. So if he wants to infer D's medical status, he has four options for disease. This is k-anonymity principle. But C knows some general details about D as his ZIP code is 14853 and age above 50. So using these values as quasi-identifiers, C concludes that D's record is present in records 5,6,7,8. But here C has three options of disease, Cancer, Heart Disease and Viral infection. Here C uses his background knowledge and concludes that D has Heart Disease because D has low blood pressure and he avoids fatty meals.

So, we can say that k-anonymity does not protect against attacks based on background knowledge. We have 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. The k-anonymity may suffer with this aspect also.

A *l* –Diversity

The protection k-anonymity provides is simple and easy to understand. If a table satisfies k-anonymity for some value k, then anyone who knows only the quasi-identifier values of one individual cannot identify the record corresponding to that individual with confidence greater than 1/k. While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute where identity disclosure means as individual is linked to a particular record in the published data and attribute disclosure means sensitive attribute information of an individual is disclosed. To address these limitations of K-anonymity, Machanavajjhalaetal. [2, 15] introduced *l*-diversity as a stronger notion of privacy.

Definition 1 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.

Fig. 4 shows 3-diverse inpatient microdata table it shows that if an adversary want any sensitive information and even he have quasi attributes value he can't access the accurate value because there are 3 more values which pretends same and this diverse factor known as diversity.

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 that we now discuss.

Observation 2: *l*-diversity may be difficult and unnecessary to achieve.

Suppose that the original data has only one sensitive attribute: the test result for a particular virus. It takes two values: positive and negative. Further suppose that there are 10000 records, with 99% of them being negative, and only 1% being positive.

	NONSENSITIVE			SENSITIVE
S. NO	ZIP CODE	AGE	NATIONALITY	MEDICAL STATUS
1	1305*	≤40	*	Heart Disease
4	1305*	≤40	*	Viral Infection
9	1305*	≤40	*	Cancer
10	1305*	≤40	*	Cancer
5	1485*	>40	*	Cancer
6	1485*	>40	*	Heart Disease
7	1485*	>40	*	Viral Infection
8	1485*	>40	*	Viral Infection
2	1306*	≤40	*	Heart Disease
3	1306*	≤40	*	Viral Infection
11	1306*	≤40	*	Cancer
12	1306*	≤40	*	Cancer

Fig. 4 3-Diverse Inpatient Microdata

Then the two values have very different degrees of sensitivity. One would not mind being known to be tested negative, because then one is the same as 99% of the population, but one would not want to be known/considered to be tested positive. In this case, 2-diversity is unnecessary for an equivalence class that contains only records that are negative. In order to have a distinct 2-diverse table, there can be at most 10000×1% = 100 equivalence classes and the information loss would be large. Also observe that because the entropy of the sensitive attribute in the overall table is very small, if one uses entropy *l*-diversity, *l* must be set to a small value.

Observation 3: *l*-diversity is insufficient to prevent attribute disclosure.

Skewness Attack: When the overall distribution is skewed, satisfying *l*-diversity does not prevent attribute disclosure. 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 records, even though the Two classes present very different levels of privacy risks.

Similarity Attack*:* When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information.[8]

Positive and Negative Disclosure: the homogeneity attack where A determined that B has Cancer is an example of Positive Disclosure, whereas when an adversary eliminates some possibilities of sensitive tuples is known as Negative Disclosure. This negative disclosure uses background knowledge attack.

l-diversity also fails in the case of multiple sensitive attributes. In short, distributions that have the same level of diversity may provide very different levels of privacy, because there are semantic relationships among the attribute values, because different values have very different levels of sensitivity, and because privacy is also affected by the relationship with the overall distribution. *l*-diversity does not consider semantic meanings of sensitive values. *l*-diversity cannot provide privacy for the multiple sensitive attributes.

b) t-Closeness

t-closeness formalizes the idea of global background knowledge by requiring that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). This effectively limits the amount of individual-specific information an observer can learn.

Intuitively, privacy is measured by the information gain of an observer. Before seeing the released table, the observer has some prior belief about the sensitive attribute value of an individual. After seeing the released table, the observer has a posterior belief. Information gain can be represented as the difference between the posterior belief and the prior belief. The novelty of our approach is that we separate the information gain into two parts: that about the whole population in the released data and that about specific individuals. To motivate our approach, let us perform the following thought experiment: First an observer has some prior belief *B*0 about an individual's sensitive attribute. Then, in a hypothetical step, the observer is given a completely generalized version of the data table where all attributes in a quasi-identifier are removed (or, equivalently, generalized to the most general values). The observer's belief is influenced by **Q**, the distribution of the sensitive attribute value in the whole table, and changes to *B*1. Finally, the observer is given the released table. By knowing the quasi-identifier values of the individual, the observer is able to identify the equivalence class that the individual's record is in, and learns the distribution **P** of sensitive attribute values in this class. The observer's belief changes to *B*2. The *l*-diversity requirement is motivated by limiting the difference between *B*0 and *B*2 (although it does so only indirectly, by requiring that **P** has a level of diversity). We choose to limit the difference between *B*1 and *B*2. In other words, we assume that **Q**, the distribution of the sensitive attribute in the overall population in the table, is public information. We do not limit the observer's information gain about the population as a whole, but limit the extent to which the observer can learn additional information about specific individuals. To justify our assumption that **Q** should be treated as public information, we observe that with generalizations, the most one can do is to generalize all quasi-identifier attributes to the most general value. Thus as long as a version of the data is to be released, a distribution **Q** will be released.1 We also argue that if one wants to release the table at all, one intends to release the distribution **Q** and this distribution is what makes data in this table useful. In other words, one wants **Q** to be public information.

A large change from *B*0 to *B*1 means that the data table contains a lot of new information, e.g., the new data table corrects some widely held belief that was wrong. In some sense, the larger the difference between *B*0 and *B*1 is, the more valuable the data is. Since the knowledge gain between *B*0 and *B*1 is about the whole population, we do not limit this gain. We limit the gain from *B*1 to *B*2 by limiting the distance Between **P** and **Q**. intuitively, if **P** = **Q**, then *B*1 and *B*2 should be the same. If **P** and **Q** are close, then *B*1 and *B*2 should be close as well, even if *B*0 may be very different from both *B*1 and *B*2.[8]

Definition 2 The t-closeness Principle: *An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.*

There are some more attacks on k-anonymity which can be controlled in some extend.

Unsorted Matching Attacks: This attack is based on the order in which tuples appear in the released table. If an adversary checks that quasi-attribute for a individual is present in any released table he can match this value with private table and fulfills his goal.. The solution of this attack is to randomly sort/substitute the values the tuples had before releasing for this sorting can be done in the basis of any attribute.

Complementary Release Attack: This is also known as linking attack. We already discussed this topic in the section 1 of this paper. Different releases can be linked together to compromise k-anonymity. For solution of this attack, other data holders may release some data that can be used in this kind of attack. Generally this kind of attack is hard to prevent completely.

Temporal Attack: Adding or removing tuples may comprom--ise k-anonymity protection. Solution of this attack is to subse--quent releases must use the already released table.

Insufficient Knowledge: The data publisher is unlikely to know the full distribution f of sensitive and on sensitive attributes over the general population from which T is a sample.

The Adversary's Knowledge is Unknown: It is also unlikely that the adversary has knowledge of the complete joint distribution between the non-sensitive and sensitive attributes. However, the data publisher does not know how much the adversary knows.

Instance-Level Knowledge: The theoretical definition does not protect against knowledge that cannot be modeled probabilistically. For example, suppose Bob's son tells Alice that Bob does not have diabetes. The theoretical definition of privacy will not be able to protect against such adversaries.

Multiple Adversaries: T here will likely be multiple adversaries with different levels of knowledge, each of which is consistent with the full joint distribution. Suppose Bob has a disease that is (a) very likely among people in the age group [30-50], but (b) is very rare for people of that age group who are doctors. An adversary who only knows the interaction of age and illness will think that it is very likely for Bob to have that disease. However, an adversary who also knows that Bob is a doctor is more likely to think that Bob does not have that disease. Thus, although additional knowledge can yield better inferences on average, there are specific instances where it does not. Thus the data publisher must take into account all possible levels of background knowledge. In the next section, we present some definitions that eliminate these drawbacks [2].

III Conclusion and Future Work

K-anonymity is used for security of respondents identity and decreases linking attack in the case of homogeneity attack a simple K-anonymity model fails and we need a concept which prevent from this attack solution is *l*-diversity. All tuples are arranged in well represented form and adversary will divert to *l* places or on *l* sensitive attributes. *l*-diversity limits in case of background knowledge attack because a no one predict knowledge level of an adversary. t-closeness helps a lot to solve this problem. But both techniques fail in the case of multiple sensitive attributes. It is observe that using generalization and suppression we also apply these techniques on those attributes which are doesn't need this extent of privacy and this lead to reduce the precision of publishing table. e- NSTAM (extended Sensitive Tuples Anonymity Method) [5] is applied on sensitive tuples only and reduces information loss, this method also fails in the case of multiple sensitive tuples. Generalization with suppression also causes of data lose because suppression emphasize on not releasing values which are not suited for K factor. Future works in this front can include defining a new privacy measure along with *l*-divesity and t-closeness for multiple sensitive attribute and.

We will focus to generalize attributes without suppression using other techniques which are used to achieve k-anonymity because suppression leads to reduce the precision of publishing table.

References

[1] V.Ciriani , S. De Capitani di Vimercati , S. Foresti ,P. Samarati,"*K*-Anonymity",Springer US, Advances In Information Security (2007).

[2] Latanya. Sweeney,"Achieving *K*-Anonymity privacy protection using generalization and suppression" International journal of Uncertainty,Fuzziness and Knowledge-based Systems,10(5), May 2002,571588.

[3] Pierangela Samarati ,Latanya Sweeney,"Protecting Privacy when Disclosing Information: *K*-Anonymity and its enforcement through Generalization and Suppression.1998.

[4] A. Machanavajjhala, J.Gehrke,D. Kifer, and M. Venkitasubramaniam.L-diversity:Privacy beyond k-anonymity. In Proc.22nd International Conf. Data Engg. (ICDE), page 24 , 2006.

[5] Xinping Hu Zhihui Sun Yingjie Wu Wenyu Hu , Jiancheng Dong " K-Anonymity Based on Sensitive Tuples", 2009 First International Workshop on Database Technology and Applications, 978-0-7695-3604-0/09 /2009 IEEE DOI 10.1109/DBTA.2009.74 M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[6] Yingjie Wu, Xiaowen Ruan,Shangbin Liao, Xiaodong Wang," P-Cover K-anonymity model for Protecting Multiple Sensitive Attributes", IEEE,The 5th International Conference onComputer Science & Education Hefei, China. August 24–27, 2010. 978-1-4244-6005-2/10/2010 IEEE.

[7] Rinku Dewri, Indrajit Ray, Indrakshi Ray ,Darrell Whitley," On the Optimal Selection of k in the k–Anonymity Problem".

[8] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. k- Anonymity: Algorithms and hardness. Technical report, Stanford University, 2004.

[9] R. J. Bayardo and R. Agrawal. Data privacy through optimal k - anonymization. In ICDE-2005, 2005

[10] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: Efficient fulldomain k-anonymity. In SIGMOD, 2005.

[11] A. Meyerson and R. Williams. On the complexity of Optimal k anonymity. In PODS, 2004.

[12] P. Samarati. Protecting respondents' identities in microdata release.In IEEE Transaction s on Knowledge and Data Engineering, 2001.

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

[14] S. Zhong, Z. Yang, and R. N. Wright. Privacy-enhancing kanonymization of customer data. In PODS, 2005

[15] A. Dobra. Statistical Tools for Disclosure Limitation in MultiwayContingency Tables. PhD thesis , Carnegie Mellon University, 2002.

[16] S. L. Kullback and R. A. Leibler. On information and sufficiency. Ann. Math. Stat., 22:79–86, 1951.