Survey on Recent Developments in Privacy Preserving Models

Sowmyarani C N
Assistant professor
Department of Computer Science and Engg,
MSRIT, Bangalore-54

Dr. G N Srinivasan
Professor
Department of Information Science and Engg,
RVCE, Bangalore-59

ABSTRACT

Privacy preserving in data mining [1] is one of the major and increasingly interested area of research under data security. Privacy will be provided for data at different levels such as, while publishing the data, at the time of retrieving result by preserving sensitive data without disclosing it. It is not just sufficient to preserve sensitive data without disclosing it, but also need to manipulate and present data so that, certain inference channels are blocked. Numbers of techniques are proposed to achieve privacy protection for sensitive data. But, most of these methods are facing side effects such as reduced utility, less accuracy, data mining efficiency down-graded, disclosure risk, etc. In this paper we analyze all these different techniques how they handle data in turn to provide privacy and points out their merits and demerits.

Keywords – *k*-anonymity, *l*-diversity, *p*-sensitive, privacy.

1. INTRODUCTION

When releasing data for analysis, preserving privacy of individual has recently raised great concern in data mining field. Organizations and agencies often need to publish micro data, for example, medical data or census data for research purpose or for giving in hands of an application developer for developing any application. There are number of techniques to preserve individual's privacy. The main concern is sensitive information should not be disclosed. There are two types of disclosures such as, identity disclosure and attribute disclosure. Identity disclosure happens when an individual can be uniquely identified from the published data. Attribute disclosure happens when the information of an individual can be inferred from the published data. The number of privacy models are discussed which are succeeded in solving the problems such as attribute disclosure and identity disclosure by preserving private information. Some of popular techniques such as K-anonymity, l-diversity and t-closeness models are discussed in this paper.

2. K-ANONYMITY

Sweeney proposed k-anonymity [2] model which assumes that person-specific data are stored as record (row) having attributes (columns) in the form of table. There are three kinds of attribute in a table such as: 1) Identifier: Attributes that uniquely identify an individual, e.g., Name, Social Security Number. 2) Quasi-identifier: Attributes whose values when taken together can potentially identify an individual, e.g., Zip-code, Birth date and Gender.

Table1. Original patterns table

	ZIP code	Age	Disease
1	54677	39	Heart Disease
2	54602	32	Heart Disease
3	54678	37	Heart Disease
4	54905	53	Gastritis
5	54909	62	Heart Disease
6	54906	57	Cancer
7	54605	40	Heart Disease
8	54673	46	Cancer
9	54607	42	Cancer

Table2. A 3-Anonymous version of Table1

	ZIP code	Age	Disease
1	546**	3*	Heart Disease
2	546**	3*	Heart Disease
3	546**	3*	Heart Disease
4	549**	>=50	Gastritis
5	549**	>=50	Heart Disease
6	549**	>=50	Cancer
7	546**	4*	Heart Disease
8	546**	4*	Cancer
9	546**	4*	Cancer

Sensitive Attribute: Attribute that is considered sensitive, e.g., Disease and Salary. It requires that each record in a table be indistinguishable from at least (k−1) other records with respect to the pre-determined quasi-identifier. K-Anonymity protects against identity disclosure, but not protects attribute disclosure which leads to homogeneity attack. Adversaries' background knowledge may lead to additional disclosure risk.

Table1 shows the original pattern. Table2 shows the anonymized version of Table1 representing 3-anonymous data. In table 2, suppose Rama knows that, Krishna is 37-years old man living at Zip code 54678, and then he can easily come to conclusion that, he is having Heart disease by disclosing identity. This constitutes homogeneity attack. Since all the sensitive attributes are homogeneous with respect to their values, It became possible to determine identity of Krishna. If Rama knows that Lava's Age and Zipcode, and he is having background knowledge that Lava is having less chances of having Heart disease, he can conclude that, Lava is having cancer. This background knowledge enables Rama to discover Lava’s identity. To address these limitations, Machanavajjhala introduced l-diversity as strong notion of privacy.

2.1 (a, k) –Anonymity Model

Definition 1 (K-anonymity): Given a table *T*(a1, a2, … , a*n*), and its quasi-identifier *Qi*, *T* satisfies *k-anonymity*[20] if and only if each sequence of values in *T[Qi]* appears with at least *k* occurrences in *T[Qi]*, where *T[Qi]* denotes the projection of attributes in *Qi*, maintaining duplicate tuples.

Definition 2 Quasi-identifier (Qi): Qi is a set of attributes in a table, which cannot identify individual by itself, but can identify individual by linking with external table.

Definition 3 (α-Deassociation): Given a dataset *D*, an attribute set *Qi* and a sensitive value *s* in the domain of attributes *S* not in *Qi*. Let (*E*, *s*) be the set of tuples in equivalence class *E* containing *s* for *S* and *α* be a user specified threshold, where 0 < *α* < 1. Dataset *D* is *α*-deassociated with respect to attribute set *Qi* and the sensitive value *s* if the relative frequency of *s* in every equivalence class is less than or equal to *α*. That is, $|(E, s)|/|E| \leq \alpha$ for all equivalence classes *E*.

Table 3. (α, k)-Anonymized table with α=0.4

Designation	Date-of-Birth	Postcode	Disease
*	1975-*-*	1541	Hep-B
*	1975-*-*	1541	Flu
*	1975-*-*	1541	Fever
*	1975-1-*	1542	Cancer
*	1975-1-*	1542	Cancer
*	1975-1-*	1542	Flu
*	1975-1-*	1542	Hep-B

Definition 4 ((α, k)-anonymity): Given an anonymity table *T* ', a quasi-identifier *Qi* and a sensitive value *s* in the domain of sensitive attribute S. *T* ' is said to be a (*α*,*k*)-anonymity [3] if *T* ' satisfies both *k*-anonymity and *α*-deassociation properties with respect to *Qi* and *s*.

A constraint *α* in simple (*α*,k) anonymity is specific to single sensitive attribute value. Deassociation value calculated based on that single value. Consider Table3 representing (*α, k)*-anonymized data with *Qi* as {Designation, Date-of-Birth, Postcode} and sensitive value s is Hep-B which satisfies (0.4,3)-anonymity. But in second equivalence class, the adversary can conclude that, for person with *Qi* there is 50% of chances that he can have Cancer. To overcome this, General (*α*,k)- Anonymity model proposed. Where, one *α* value is set for all sensitive attributes in the equivalence class. Problem with general (*α*,k)- anonymity is it sets uniform value for *α* on all sensitive values, But each sensitive value will have different level of sensitivity.

Table 4. General (*α*, k) - Anonymized table with *α*=0.4

Designation	Date-of-Birth	Postcode	Disease
*	1975-*-*	154*	Hep-B
*	1975-*-*	154*	Flu
*	1975-*-*	154*	Fever
*	1975-1-*	154*	Cancer
*	1975-1-*	1542	Cancer
*	1975-1-*	1542	Flu
*	1975-1-*	1542	Hep-B

Table 4 representing general (*α*, k)-anonymized data with uniform value of α for all sensitive values. But in the table, disease column values are have different level of sensitive values. Considering Hep-B, Flu and Fever are less sensitive compared to Hep-B.

2.2 *p*-Sensitive, *K*-Anonymity Model

This model is proposed to overcome k-anonymity attacks. Table 5 shows the 2-anonymous data with quasi-identifier set *Qi* with attribute set {Age, Country, Zipcode} and sensitive attribute s is disease. Depending on the sensitivity of sensitive attribute disease, the values can be categorized into 4 categories as shown in the table 6. Even though the Table 5 is 2-anonymous, for Richa and Raman we can conclude that, they have Cancer, by referring to Table 6 which shows the background knowledge available for them.

Table 5. 2-Anonymous table view of micro data

ID	Age	Country	Zipcode	Disease
1	<30	America	152**	HIV
2	<30	America	152**	HIV
3	<30	America	1524*	Cancer
4	<30	America	1524*	Cancer
5	>40	Asia	120**	Hepatitis
6	>40	Asia	120**	Phithisis
7	>40	Asia	120**	Asthma
8	>40	Asia	120**	Heart Disease
9	3*	America	1524*	Fever
10	3*	America	152**	Fever
11	3*	America	152**	Fever
12	3*	America	1524*	Gastritis

To overcome this, *p*-sensitive *k*-anonymity[4] concept proposed.

1.2.1 Definition 5. (p-sensitive k-anonymity): The table *T* satisfies *p*-sensitive *k*- anonymity property if it satisfies *k*-anonymity, and for each equivalence class *Qi* in *T*, the number of distinct values for each sensitive attribute occurs at least *p* times within the same equivalence class *Qi*.

2.3 (*p*+, *α)*-sensitive k-anonymity

Sometimes, the sensitive values are categorized into different parts as shown in Table 7. In this table, disease is a sensitive attribute, categorized according to the sensitivity of the disease.In such case, data owner is interested to preserve top secret information such as diseases like HIV, Hepatitis in this example. In the first *Qi* group, sensitive values such as {HIV, HIV, Cancer, Cancer}are distinct but, belongs to same top secret category. To avoid such situations, another approach is proposed, called as (*p*+, *α)*-sensitive k-anonymity [5].

Definition 6. *((p+, α)-**sensitive** k-**anonymity**):* The table T satisfies (p+, α)-sensitive k-anonymity [5] property if it satisfies k-anonymity, and each Qi-group has at least p distinct categories of the sensitive attribute and its total weight is at least α.

Table 5 can be represented as Table 9 by applying *(p+, α)*-sensitive k-anonymity by replacing the categories as values of sensitive attributes.

Table 6. External Information available

Name	Age	Country	Zipcode
Richa	26	USA	15246
Nick	45	India	13064
Raman	25	Canada	15249
Yih jyh	48	Japan	13074

Table 7. Categories of Disease

Category_ID	Sensitive attribute values	Sensitivity
One	HIV, Cancer	Top Secret
Two	Phthisis, Hepatitis	Secret
Three	Heart Disease, Asthma	Less Secret
Four	Fever, Gastritis	Non Secret

Table 8. 2-sensitive 4-anonymous data

ID	Age	Country	Zipcode	Disease
1	<30	America	152**	HIV
2	<30	America	152**	HIV
3	<30	America	152**	Cancer
4	<30	America	152**	Cancer
5	>40	Asia	120**	Hepatitis
6	>40	Asia	120**	Phithisis
7	>40	Asia	120**	Asthma
8	>40	Asia	120**	Heart Disease
9	3*	America	152**	Fever
10	3*	America	152**	Fever
11	3*	America	152**	Fever
12	3*	America	152**	Gastritis

Table 9. (2^+, 2)-sensitive 4-anonymous data

ID	Age	Country	Zipcode	Disease
1	<40	America	1524*	HIV
2	<40	America	1524*	Cancer
3	<40	America	1524*	Fever
4	<40	America	1524*	Gastritis
5	>40	Asia	120**	Heppatitis
6	>40	Asia	120**	Phithisis
7	>40	Asia	120**	Asthama
8	>40	Asia	120**	Heart Disease
9	<40	America	1520*	HIV
10	<40	America	1520*	Cancer
11	<40	America	1520*	Fever
12	<40	America	1520*	Fever

Another approach called p-cover k-anonymity [6, 22] proposed to overcome the limitations of p-sensitive k-anonymity. This solves the problem of multiple sensitive attribute disclosure, by providing high quality of data. Extra associations among multiple sensitive attributes will be investigated.

3. L-DIVERSITY

The notion of l-diversity [7] attempts to solve a problem by requiring that, each equivalence class should have at least l-well represented values for each sensitive attributes.

Table 10. Original pattern

ID	Age	Zipcode	Disease
1	26	15246	Stomach Ulcer
2	25	13064	Toothache
3	45	14249	Gastritis
4	48	14274	Gastritis

Definition 7 l-diversity: A table is said to have *l*-diversity if every equivalence class should have are at least *l* well-represented values for the sensitive attribute.

Table 11. 2-diversity micro data

ID	Age	Zipcode	Disease
1	2*	150**	Stomach Ulcer
2	2*	150**	Toothache
3	4*	142**	Gastritis
4	4*	142**	Stomach Ulcer

When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information. In Table 11, consider that, an intruder knows Arun's age is around 20 and knows his zip code. Even though the values are diverse, stomach ulcer and gastritis are stomach related disease. So, he can come to conclusion that Arun is having stomach related disease. To overcome such problems, another approach proposed called as (*a*, *d*)-Diversity.

3.1 (*a*, *d*)-Diversity

The real meaning of the sensitive attribute values into consideration, the principle of (*a*, *d*)- diversity [8] is derived as:

Table 12. Micro data with (2, 2)-diversity

ID	Age	Zipcode	Disease
1	2*	150**	Stomach Ulcer
2	2*	150**	Gastritis
3	2*	150**	Bronchitis
4	2*	150**	Asthma
5	4*	142**	Gastritis
6	4*	142**	Stomach cancer
7	4*	142**	Asthma
8	4*	142**	Bronchitis

Definition 8. (a, d)- diversity: An equivalent class is said to satisfy (a, d)-diversity if it contains at least a analogous values for sensitive attributes S, and at least d dissimilar values for S. A table is said to satisfy (a, d)- diversity if every equivalent class satisfies (a, d)- diversity.

Analogous Values: Given certain characteristic, two values are analogous if they have the most in common.

Dissimilar Values: Given certain characteristic, two values are dissimilar, if they have nothing or little in common.

Table 12 shows (2, 2)-diversity. In the table, its difficult to determine Arun's disease, since the gastritis and stomach ulcer are stomach related disease and bronchitis, asthma are lungs diseases.

3.2 Unique Distinct l-SR Diversity

Consider the example of knowing about the person that, he is having fever is totally different from knowing the person is having HIV with respect to impact. With the knowledge of what is socially acceptable, we can obtain the level of sensitivity of the information. This approach is proposed challenging distinct l-diversity model which does not prevent probabilistic inference attack.

3.2.1 *Sensitivity of Private Information:* In l-SR Diversity[22], sensitivity of private information is considered as important concern. Private information refers to individually identifiable information and sensitivity of private information refers to the impact of disclosure of that information.

3.2.2 *Sensitivity Attack:* Sensitivity attack [9] happens when the sensitivity level of sensitive attributes in one equivalence class falls into a narrow range, so the adversary can learn the sensitivity of such information.

3.2.3 *Sensitivity Ranking (SR):* The diversity on sensitivity levels of sensitive attributes is considered in sensitivity ranking [9]. The idea is to rank distinct values of sensitive attributes and represent them as sensitivity ranking levels. The higher the sensitivity ranking level, there is more the impact of data disclosure.

Definition 9. Unique Distinct l-SR diversity[9]: A table is said to satisfy Unique Distinct l- SR diversity if each of its equivalence class contains exactly l distinct sensitivity ranking levels.

4. T-CLOSENESS

In this approach, privacy can be measured in terms of information gained by the observer. An observer can gain information based on prior belief before seeing the released data and post belief after seeing the released data. So, the information gain is can be represented as the difference between the posterior belief and the prior belief. The approach separates the information gain into two parts: about the whole population in released data and about the specific individuals.

Consider B0 as a prior belief of an observer and B1 is belief of an observer changed after seeing the generalized data. The observer gains some more information by knowing quasi-identifier values in equivalence class and changes his belief to B2 based on assuming to which equivalence class the individual belongs to. Assume that, Q is the distribution of the sensitive attribute in the overall population in the table. Observer learns a distribution P by knowing the quasi-identifier values of the individual. So that he is able to identify the equivalence class to which the individual's record belongs to. The approach limits the gain from B1 to B2 by limiting the distance between P and Q. But not limit the gain between B0 and B1, because it is about the whole population. Intuitively, if P = Q, then B1 and B2 should be the same. If P and Q are close, then B1 and B2 should be close as well, even if B0 may be very different from both B1 and B2.

Definition 10. t-closeness: 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.

5. COMPARISONS

In the following table, there are different types of attacks mentioned with respect to privacy model. The similarity and homogeneous attacks seems same, but the homogenous means, the sensitive attribute values will be one and the same.

Table 13. Comparison of privacy models considering attacks.

Privacy Models	Attack Models				
	Homogeneity attack	Skewness attack	Membership disclosure attack	Similarity attack	Probabilistic inference attack
k-anonymity	✓				
(a, k) – Anonymity Model	✓				
p-Sensitive, k-anonymity Model			✓		
(p+, α)-sensitive k-anonymity			✓		
l-diversity		✓		✓	
(*a*, *d*)-Diversity				✓	
Unique Distinct l-SR Diversity					✓
t-closeness		✓		✓	

In similarity attack, values of sensitive attributes will be only semantically same. The different implementations of diversity models experience similarity attacks [11] when the values of sensitive values are semantically same and probabilistic inference attacks when observer can infer the sensitive values based on background knowledge. Anonymity models experience homogeneity and membership disclosure attacks. t-closeness experiences skewness attack [11] and similarity attack.

6. CONCLUSION

Person-specific data in its original form often contains sensitive information. Since there are lots of opportunities for adversaries to disclose sensitive data, it cannot be directly published or released. So, several privacy models are developed recently to preserve private information. In this paper, number of privacy preserving models are discussed and compared with respect to the attacks they experience. There are many challenges in this research field such as: Techniques should limit the disclosure risks while minimizing the utility of the data and to balance privacy and accuracy level of data. Privacy preserving technology needs to be further researched to minimize these complexities of the privacy problem. The further developments of these privacy models in this direction would leads to an added advantage to solve these complexities to preserve the privacy.

7. REFERENCES

[1] Agrawal, R. and Srikant, R, “Privacy-preserving data mining”, In Proc. SIGMOD00, 2000, pp. 439-450.

[2] L. Sweeney, “*k*-anonymity: a model for protecting privacy”, International Journal on Uncertainty, Fuzziness and Knowledge based Systems, 2002, pp. 557-570.

[3] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang. (alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *KDD*, pages 754–759, 2006.

[4] TRUTA T M , VINAY B, “Privacy protection: *p*-Sensitive *k*-anonymity property,” Proceedings of the 22nd on Data Engineering Workshops, IEEE Computer Society, Washington Dc, 2006.

[5] Xiaoxun Sun, Hua Wang, Jiuyong Li, Truta, T.M, Ping Li, “(p+, α) -sensitive k-anonymity: A new enhanced privacy protection model” , 2008,pp.59-64.

[6] Yingjie Wu, Xiaowen Ruan, Shangbin Liao, Xiaodong Wang, “P-cover k-anonymity model for protecting multiple sensitive attributes”, 2010,pp.179-183.

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

[8] Qian Wang, Xiangling Shi,” (a, d) Diversity: Privacy Protection Based on l-Diversity” Software Engineering, 2009 pp.367-372.

[9] Yunli Wang, Yan Cui, Liqiang Geng, Hongyu Liu, “A new perspective of privacy protection: Unique distinct l-SR diversity”, Privacy Security and Trust (PST), 2010 Eighth Annual International Conference 2010 PST pp.110-117.

[10] A. Meyerson and R. Williams. “On the complexity of optimal kanonymity [C]”. In: Proc of the ACM SIGMOD Int’l Conf on Principles of DB Systems. New York: ACM Press, 2004. 223-228.

[11] Domingo-Ferrer, J, Torra, V,” A Critique of k-Anonymity and Some of Its Enhancements” 2008, pp.990-993.

[12] Agrawal, R. and Srikant, R, “Privacy-preserving data mining”, In Proc. SIGMOD00, 2000, pp. 439-450.

[13] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, “State-of-the-art in privacy preserving dat mining”, In Proc of ACM SIGMOD, 2004, pp. 50–57.

[14] P. Samarati. Protecting respondents’ identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027, 2001.

[15] T. M. Truta and B. Vinay. Privacy protection: p-sensitive k-anonymity property. In 2nd International Workshop on Privacy Data Management PDM 2006, page p. 94, Berlin Heidelberg, 2006. IEEE Computer Society.

[16] L. Willenborg and T. DeWaal. Elements of Statistical Disclosure Control. Springer-Verlag, New York, 2001.

[17] R. C.-W. Wong, A. W.-C. Fu, Ke Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In Proceedings of the VLDB 2007, Vienna, 2007.

[18] WONG R C, LI J, FU A W, et a1, “(α, k)-Anonymity : an enhanced k-anonymity model for privacy-preserving data publishing”, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, New York, 2006, pp. 754-759.

[19] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte- Nordholt, G. Seri, and P.-P. DeWolf. Handbook on Statistical Disclosure Control (version 1.0). Eurostat (CENEX SDC Project Deliverable), 2006.

[20] Vijayarani,S, Tamilarasi, A. Sampoorna, M. “Analysis of Privacy Preserving K-Anonymity Methods and Techniques” Proceedings of the International Conference on Communication and Computational Intelligence – 2010, Kongu Engineering College, Perundurai, Erode, T.N.,India.27 – 29 December,2010.pp.540-545.

[21] Wang, Y., Cui, Y., Geng, L., and Liu, H. A new perspective of privacy protection: Unique distinct l-SR diversity. In Proceedings of PST. 2010, 110-117.

[22] Yingjie Wu; Xiaowen Ruan; Shangbin Liao; Xiaodong Wang; “P-Cover K-anonymity model for Protecting Multiple Sensitive Attributes”The 5th International Conference on Computer Science & Education Hefei, China. August 24–27, 2010