Chapter 99
Research on Privacy Preserving Based on K-Anonymity

Xiao-ling Zhu and Ting-gui Chen

Abstract K-anonymity is a highlighted topic of privacy preservation research in recent years, for it can effectively prevent privacy leaks caused by link attacks; so far K-anonymity has been widely used in all fields. In this chapter, based on the existing K-anonymity privacy protection of the basic ideas and concepts, K-anonymity model, and enhanced the K-anonymity model has been studied, finally, the future directions in this field are discussed.

Keywords Privacy preservation · K-anonymity · Generalization and suppression · The enhanced K-anonymity

99.1 Introduction

With the rapid development of information technology and networks, the emergence of data mining make it possible for people to get useful information from a large database, data mining has been widely used in retail, health care, education, insurance, banking and other fields. Data Release, as an effective means of information exchange and data sharing, has brought great convenience for data retrieval and use. However, the leakage of sensitive information are also increasingly prominent

X. Zhu
College of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, Zhejiang, China
e-mail: lynn_525@sina.com

T. Chen (✉)
Contemporary Business and Trade Research Center, Zhejiang Gongshang University, Hangzhou 310018, Zhejiang, China
e-mail: ctgsimon@gmail.com

X. He et al. (eds.), *Computer, Informatics, Cybernetics and Applications*, Lecture Notes in Electrical Engineering 107, DOI: 10.1007/978-94-007-1839-5_99,

during the data released, data Mining providing people with a strong knowledge discovery, while it also posing a threat of personal privacy, privacy protection has become a hot topic in database security research. A reasonable and effective method of protection, which can protect the user's privacy and keep the data available, is the trend of developments in information security.

The existing data mining methods are: heuristic-based privacy protection technology, cryptography-based privacy preserving techniques, and privacy protection technology based on the reconstruction, for different methods they applied well in related fields, and can protect user's privacy information to some extent. Now the commonly used method is the K-anonymity, aiming at the existing K-anonymity to summarize the main ideas and models and analysis, and the future direction of development are discussed.

99.2 K-Anonymity Model

In real life, some agencies often should publish some relevant data, such as for the need of research on population statistics, medical and health. Although the published data has been hidden personal identifiable information, such as name, ID number, telephone number and other attributes. The attacker get data through other channels to obtain the data link operations to infer the privacy of data, resulting in privacy leak, this process is called link attacks, which mainly work in quasi-identifier. Currently, to avoid link attacks, and protecting private information from being leaked, K-anonymous method is the most common.

K-anonymity was raised in 1998 by Samarati and Sweeney, it requires the published data exists a certain number (at least for the K) whose records cannot be distinguished, so that an attacker cannot distinguish the respective privacy information of a specific individuals, thereby it prevents the leakage of personal privacy. User can specify a parameter K for the greatest risk of information leakage that they can receive in K-anonymous. It protects the privacy of individuals to some extent, while it also reduces the availability of data; the work of K-anonymity focuses mainly on the protection of private information and increase their availability. Since the proposed, K-anonymity has been the general concern of academia; many scholars at home and abroad research and develop the technology in different way (Fig. 99.1).

99.3 K-Anonymity Model for the Main Algorithm

99.3.1 Generalization and Suppression

The aim of the current data mining is focused on how to set the original data by anonymous effectively data, and at the same time to achieve the best anonymity, the maximum data availability, the minimum spending of time and space.

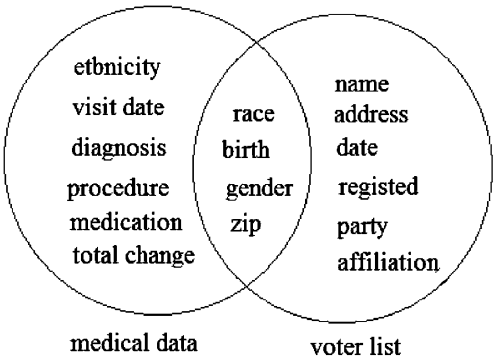

Fig. 99.1 Link attacks

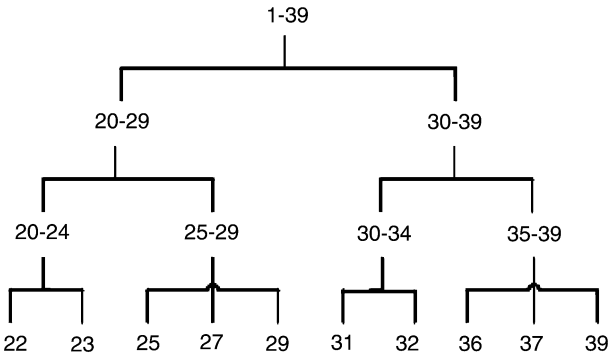

Fig. 99.2 Generalization of age

Different with distortion of data interference methods such as distortion, disturbance and randomization, K-anonymity can maintain the authenticity of the data. The common use to achieve K-anonymity is through generalization and suppression. Generalization can be divided into value generalization and domain generalization, Fig. 99.2 describes the generalization of age. Suppression is to delete or hide some of the attributes in the data table directly to protect patient privacy.

Table 99.1 shows a medical information form, after inhibition, it does not include sensitive information such as their names, Medicare numbers, home address, and ID numbers and so on. But there exist quasi-identifier information such as gender, age, zip code, etc. Through the collection of these attributes, attackers can tell the personal information indirectly, so it is possible to disclose patient medical information.

To prevent the disclosure of private information and to protect patient privacy, Table 99.2 generalized the data, the patient information in the table specified the age in a large range, zip code in the "*" indicates any number, the rationale is as follows in Table 99.2.

Table 99.1 Original medical information

ID	Quasi-identifier			Other	Sensitive information
	Gender	Age	Zip		
1	Male	25	644625	……	Gastric ulcer
2	Male	27	650500	……	AIDS
3	Male	22	671060	……	Flu
4	Male	29	675067	……	Neurasthenia
5	Female	34	671060	……	Flu
6	Female	31	671060	……	Hepatitis
7	Female	37	650500	……	Neurasthenia
8	Female	36	650500	……	Flu

99.3.2 The Enhanced K-Anonymity Model

The original K-anonymity is to prevent identity disclosure, but through the property will still bring the disclosure of information. In order to solve the shortcomings in K-anonymity, Machanavajjhala proposed K-anonymity model for the two attack methods, homogeneity attack and the background knowledge attack. Homogeneity attack is the attacker derived K-anonymous table information of a sensitive individual; background knowledge inference attacks is that attackers use some additional information in advance to carry out attacks. The two attacks will result in disclosure of sensitive property in K-anonymity.

99.3.2.1 L-Diversity Model

Machanavajjhala, who gives L-diversity model to reduce the privacy, leaks by increasing the diversity of an anonymous in group of sensitive properties. In a published table, a K-anonymous group contains at least L sensitive properties that on behalf of a good sense of representative. For example, in Table 99.2, tuples whose ID number are 1, 2, 3, 4, 6 form a group with 5 species diversity, their frequency was 12.5, 12.5, 37.5, 25 and 12.5% in value, and no one has predominant function, so it can be set by L-diversity model.

However, in this model, it is difficult to speculate how much background knowledge the attacker knows about, any posted data will become unsafe if the other knows a lot about the patient's background knowledge, there does not have a good way to set the parameters in L-diversity model.

99.3.2.2 (α, k)-Anonymous Model

In (α, k)-anonymous model, property with higher degree of sensitivity has been better protected, by constraining the frequency of anonymity property values in the

Table 99.2 Generalization of the data

ID	Quasi-identifier			Other	Sensitive information
	Gender	Age	Zip		
1	Male	25	644***	……	Gastric ulcer
2	Male	27	650***	……	AIDS
3	Male	22	671***	……	Flu
4	Male	29	675***	……	Neurasthenia
5	Female	34	671***	……	Flu
6	Female	31	671***	……	Hepatitis
7	Female	37	650***	……	Neurasthenia
8	Female	36	650***	……	Flu

sensitive group less than a given parameter α, so it avoids the situation that the frequency of certain sensitive information too high, increases the diversity of sensitive values, and prevents the consistency attack. For example, in a data sheet of a hospital medical records, some patients' illness are more sensitive and need protection, such as AIDS; while there are many diseases that are very common and not need to protect, such as flu, so under such a circumstance that a data sheet with different protection needs that the higher ones need to be protected, a more suitable idea comes out, that is (α, k)-anonymous mode.

In (α, k)-anonymous model, only the relevant and sensitive property values are necessary to protect, so only consider the sensitive attribute value, such as AIDS. This model allows inference between credibility to the sensitive lower than α, it is simple and effective way to prevent sensitive to the value for the homogeneity attack (Table 99.3).

Table 99.3 provides a (α, k)-anonymous form. In the table, flu and neurasthenia is not considered as sensitive information, from the (female 30–39, 671***) to the reasoning of depressive neurosis credibility value is 25%.

The attackers could not see the value of property of higher degree of sensitivity after processing by (α, k)-anonymous model, so it can protect the security of this kind of information effectively.

While (α, k)-anonymous model only considers the sensitive attribute the highest level-sensitive property value, there is no other level of sensitive property to process property values, and does not take into account the sensitivity of the same property value, so the same level of sensitive attributes property values or the existence of other levels of privacy disclosure.

99.3.2.3 (α, L)-Diversification K-Anonymity Model

(α, L)-anonymity model considers only the highest level of the sensitive property value, but neither other level of sensitive property values, nor take into account the sensitivity property values of the same issues. And sometimes it is hard to

Table 99.3 (0.25, 3)-Anonymous

ID	Quasi-identifier			Other	Sensitive information
	Gender	Age	Zip		
1	Male	25	644***	……	Flu
2	Male	27	650***	……	Neurasthenia
3	Male	22	671***	……	Flu
4	Male	29	675***	……	Neurasthenia
5	Female	34	671***	……	Flu
6	Female	31	671***	……	Flu
7	Female	37	650***	……	Neurasthenia
8	Female	36	650***	……	Flu

Table 99.4 Health categories

ID	Value	Sid
ID	Value	Sid
1	AIDS, hepatitis	1
2	Gastric ulcer	2
3	Neurasthenia, flu	3

determine whether an illness is the higher degree of sensitivity that needs to be protected, or lower degree of sensitivity that not need to be protected

K-anonymity model in (α, L)-diversification K-anonymity model can determine flexibility to protect the privacy or not according to the extent of protection. At the same time, have special treatment to the high-level sensitive property values of privacy protection, and with better privacy protect effect. Table 99.4 is a Health categories table (where Sid is sensitive to the privacy level of property value) is Sid the data in Table 99.2 for the classification.

(α, L)-diversification K-anonymity model is a data table to meet K-anonymity, α-distribution, and the number of Sid in the sensitive group is no less than L at the same time. α-distribution constraint is that all of the sensitive attributes of privacy frequency of Sid from equivalent class less than α which is a given data, that is sensitive to all the anonymous group of private property when the degree of the frequency Sid $\leq \alpha$, where a is user-defined number, and $0 < \alpha < 1$.

Assuming the Sid need to protect equals to 1, privacy protection level is as classification in Table 99.4, while Table 99.5 is a constraint to meet the 0.5 distribution of a data set. In Table 99.5, there are two anonymous group: {1, 2, 3, 4} and {5, 6, 7, 8}, in the first anonymous group, the frequency is 0.25 when Sid equals to 1, and in the second one the frequency is 0.25, so for all anonymous groups when Sid equals to 1 the frequency of Sid $\leq$ 0.25. Then it meets to (0.25, 3)-diversification 4-anonymous as shown in Table 99.5.

Table 99.5 provides a data table satisfying (0.25, 3)-diversification 4-anonymous model, according to the foregoing, the distribution of this data sheet meet to the constraints of 0.25, and the number of each anonymous tuple is no less than 4,

Table 99.5 (0.25, 3)-Diversification 4-anonymous

ID	Quasi-identifier			Other	Sensitive information
	Gender	Age	Zip		
1	Male	25	644***	……	2
2	Male	27	650***	……	1
3	Male	22	671***	……	3
4	Male	29	675***	……	3
5	Female	34	671***	……	3
6	Female	31	671***	……	1
7	Female	37	650***	……	3
8	Female	36	650***	……	3

the different number of the Sid values in the table equals to 3, so Table 99.4 satisfy(0.25, 3)-diversification 4-anonymous model.

Construct (α, L)-diversification K-anonymity model algorithm as follow:

Input: data set T;

Output: data table T * that meet (α, L)-diversification K-anonymity model.

(1) According to the health status in Table 99.5, the value of the sensitive property in table T was replaced by Sid, who represent the sensitive level, then table T turns into table T1.
(2) Construct a data table T2 that consistent with (α, K) data tables anonymous model, in which Sid is regarded as the sensitive property, and the generalization in accordance with top–down algorithm.
(3) For each anonymous group, check the privacy level Sid for the number of different values.
(4) IF (3L).
(5) Return the final table T*.
(6) Else
(7) Of all the anonymous groups that does not meet the requirement, have them further generalization or exchange tuples to make sure that the value of Sid is greater than L.
(8) Returns the final data table T*.

First, the value of the sensitive property in table T was replaced by Sid who represent the sensitive level, according to the health status in Table 99.5, then table T turns into table T1, and then turn T1 into anonymity, so as to meet K-anonymous and α-distribution, in this step, top–down local generalization algorithm has been used. And then check whether the generalization of the data sheet meets (α, L)-diversification K-anonymity model conditions, that is the privacy degree of different values greater than or equal to L. If all anonymous groups met for the condition, the entire data table is the final meet (α, L)-diversification K-anonymity model data tables, and if not, have further generalization or suppression to make sure that the different values of privacy degree number is greater than L.

In (α, L)-diversification K-anonymity model, the parameter α can be set by users themselves according to their privacy protection needs. It provides an effective solution to the problem of imbalance distribution of sensitive attributes, divides the attribute values on the sensitive level of privacy protection, and protects the privacy effectively.

The enhanced K-anonymity models are mainly based on K-anonymity and to make the information security. L-diversity model in which properties are divided into groups, by increasing the variety in groups, it can prevent attackers from locating the information; (α, k)-anonymous model, by processing to the higher level of sensitive attribute and make its feasible degree smaller than α, can effectively protect sensitive information of higher degree; (α, L)-diversification K-anonymity model divides the properties according to the level of sensitive information which is determined by the users themselves flexibility, and for the sensitive attributes with higher degree value require special treatment.

99.4 Development Trend and Summary

Because K-anonymity can prevent users' private information from being leaked in the released environment, ensure the authenticity of the published data, it applications widely in the industry and attracted widespread attention. However, nowadays the majority of K-anonymity algorithms are based on static data sets, and in the real world, data is constantly changing, including changes in forms of data, attribute changes, adding new data, and deleting the old data. Besides, the data between data sets are likely to be interrelated, how to achieve privacy protection in a much more complex environment with dynamic data, still need further study.

Acknowledgments This research is supported by Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20103326120001), Zhejiang Provincial Natural Science Foundation of China (No. Y7100673), Zhejiang Provincial Social Science Foundation of China (Grant No. 10JDSM03YB), and the Contemporary Business and Trade Research Center of Zhejiang Gongshang University (No. 1130KUSM09013 and 1130KU110021) as well as Research Project of Department of Education of Zhejiang Province (No. Y200907458). We also gratefully acknowledge the support of Science and Technology Innovative project (No. 1130XJ1710215).

References

1. Sweeney L (2002) Achieving k-anonymity privacy protection using generalization and suppression [J]. Intern J Uncertain Fuzziness Knowl-based Syst 10(5):571–588
2. Sweeney L (2002) k-Anonymity: a model for protecting privacy [J]. Intern J Uncertain Fuzziness Knowl-based Syst 10(7):557–570
3. Sweeney L (2001) Computational disclosure control: a primer on data privacy protection, Ph. D. thesis Massachusetts Institute of Technology, pp 67–82

4. Samarati P, Sweeney L (1998) Generalizing data to provide anonymity when disclosing information (abstract). In: Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems, Seattle, p 188
5. Long Q (2010) Privacy protection based on k-anonymity [J]. Sci Technol Assoc Forum 3:41–43
6. Qin X, Men A, Zou Y (2010) Privacy protection based on Kanonymity algorithms [J]. J Chifeng Univ 26(5):14–16
7. Wang P, Wang J (2010) Progress of research on K-anonymity privacy2preserving techniques. J Chi-feng Univ (Nat Sci Ed) [J] 27(6):2016–2019
8. Kan Y, Cao T (2010) Enhanced privacy preserving K-anonymity model (α, L)-diversity K-anonymity. Comput Eng Appl 46(21):148–151
9. Cen T, Han J, Wang J, Li X (2008) Survey of K-anonymity research on privacy preservation. Comput Eng Appl 44(4):130–134
10. Wong R, Li J, Fu A et al (2006) (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing [C]. In: International conference on knowledge discovery and data mining, vol 123. pp 754–759
11. Aggarwal G, Feder T, Kenthapadi K et al (2005) Anonymizing tables[C]. In: Proceedings of the 10th international conference on database theory (ICDT05), Edinburgh, Scotland, pp 246–258
12. Machanavajjhala A, Gehrke J, Kifer D (2007) l-Diversity: privacy beyond k-anonymity [J]. ACM Trans Knowl Discov Data 1(1):24–35
13. Li N, Li T, Venkatasubramanian S (2007) T-closeness: privacy beyond k-anonymity and l-diversity [C]. ICDE, pp 106–115
14. Fung B, Wang K, Yu P (2005) Top–down specialization for information and privacy preservation [C]. In: Proceedings of the 21st international conference on data engineering (ICDE05), Tokyo, Japan
15. Meyerson A, Williams R (2004) On the complexity of optimal K-anonymity [C]. In: Proceedings of the 23rd ACM-SIGMOD-SIGACTSIGART symposium on the principles of database systems, Paris, France, pp 223–228
16. LeFevre K, DeWitt DJ, Ramakrishnan R (2005) Incognito: efficient full-domain k-anonymity. In: SIGMOD' 05 proceedings of the 2005 ACM SIGMOD international conference on management of data, pp 49–60
17. Bayardo R, Agrawal R (2005) Data privacy through optimal K-anonymity. In: ICDE05: the 21st international conference on data engineering, pp 217–228