Implementation of Privacy Algorithm for Data Publishing

S. Subashini[1] R. Mohana Bharathi[2]
[1]M. E. (Final Year) [2]Assistant Professor
[1, 2]Selvam College of Technology, Namakkal.

Abstract---With numerous organizations collecting customer data there exists a possibility of data sharing for exploring interesting data about behavior of customers. This leads to identification of customers which can be treated as a privacy threat according to directives. These acts insist that anonymity should be guaranteed if the customers wish so. A customer data normally contains attributes like SSN name, age, postal code, date of birth and gender. This data enables identification of the individuals even though information like SSN and Name suppressed. This was first identified. The solution proposed k-anonymity property to be applied to the data before release. Subsequently several solutions were published. Most of them addressed issues related to preserving privacy of individuals related to a single organization. This paper discusses an approach to protect privacy when anonymized data of two or more organizations is integrated.

I. Techniques Used

A. *Data anonymization*

While the released table gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the table. Therefore, our objective is to limit the disclosure risk to an acceptable level while maximizing the benefit.

This is achieved by anonymizing the data before release. The first step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge. For example, knowing a particular individual in person, or from other publicly available databases like a voter registration list that include both explicit identifiers and quasi-identifiers.

A common anonymization approach is generalization, which replaces quasi-identifier values with values that are less-specific but semantically consistent. As a result, more records will have the same set of quasi-identifier values. We define an equivalence class of an anonymized table to be a set of records that have the same values for the quasi-identifiers.

To effectively limit disclosure, we need to measure the disclosure risk of an anonymized table. To this end, introduced k-anonymity as the property that each record is indistinguishable with at least k-1 other records with respect to the quasi-identifier. In other words, k-anonymity requires that each equivalence class contains at least k records.

B. *Disclosure identification*

Government agencies and other organizations often need to publish microdata, for example, medical data or census data, for research and other purposes. Typically, such data are stored in a table, and each record or row corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories:

1) Attributes that clearly identify individuals. These are known as explicit identifiers and include, e.g., Social Security Number.
2) Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers, and may include, e.g., Zip code, Birth-date, and Gender.
3) Attributes that are considered sensitive, such as Disease and Salary. When releasing micro data, it is necessary to prevent the sensitive information of the individuals from being disclosed.

Two types of information disclosure have been identified in the literature namely Identity disclosure and Attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when new information about some individuals is revealed, i.e., the released data make it possible to infer the characteristics of an individual more accurately than it would be possible before the data release.

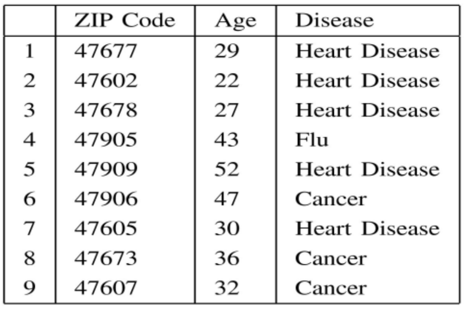
Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is reidentified and the corresponding sensitive values are revealed. Attribute disclosure can occur with or without identity disclosure.

It has been recognized that even disclosure of false attribute information may cause harm. An observer of a released table may incorrectly perceive that an individual's sensitive attribute takes a particular value and behaves accordingly based on the perception. This can harm the individual, even if the perception is incorrect.

C. *Data protection*

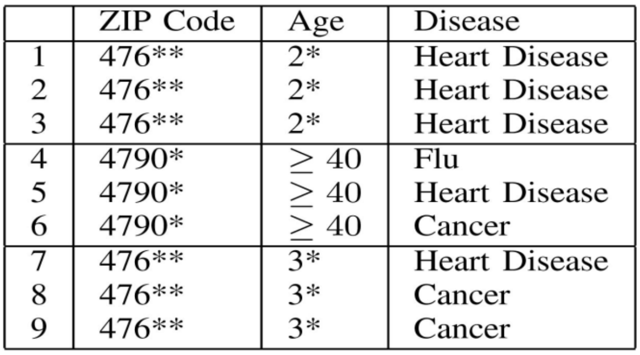
The protection k-anonymity provides is simple and easy to understand. If a table satisfies k-anonymity for some value k, then anyone who knows only the quasi-identifier values of one individual cannot identify the record corresponding to that individual with confidence greater than 1=k.

While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure.



	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Table. 1: Original data table.



	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Table. 2: Anonymized version of data table.

Let consider, Table 1 is the original data table, and Table 2 is an anonymized version of it satisfying k-anonymity. The Disease attribute is sensitive. Suppose Alice knows that Bob is a 27 year old man living in ZIP 47678 and Bob's record is in the table.

From Table 2, Alice can conclude that Bob corresponds to one of the first three records, and thus, must have heart disease. This is the homogeneity attack. For an example of the background knowledge attack, suppose that by knowing Carl's age and zip code, Alice can conclude that Carl corresponds to a record in the last equivalence class in Table 2. Furthermore, suppose that Alice knows that Carl has a very low risk for heart disease. This background knowledge enables Alice to conclude that Carl most likely has cancer.

D. Privacy measure

In this paper, the proposed novel privacy notion called "closeness."We first formalize the idea of global background knowledge and propose the base model t-closeness which requires that the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table.

This effectively limits the amount of individual-specific information an observer can learn. However, an analysis on data utility shows that t-closeness substantially limits the amount of useful information that can be extracted from the released data.

This limits the amount of sensitive information about individuals while preserves features and patterns about large groups. To incorporate distances between values of sensitive attributes, Earth Mover Distance metric is used to measure the distance between the two distributions. Then also show that EMD has its limitations and describe our desired data for designing the distance measure.

Then also, the proposed novel distance measure that satisfies all the requirements. Finally, we evaluate the effectiveness of the closeness model in both privacy protection and utility preservation through experiments on a real data set.

E. Data publishing

Privacy-preserving data publishing has been extensively studied in several other aspects.

First, background knowledge presents additional challenges in defining privacy requirements.

Second, several works considered continual data publishing, i.e., republication of the data after it has been updated. It presence to prevent membership disclosure, which is different from identity/attribute disclosure. showed that knowledge of the anonymization algorithm for data publishing can leak extra sensitive information.

References

[1] S. Goryczka, L. Xiong, and B. C. M. Fung, "m-Privacy for collaborative data publishing," in Proc. of the 7th Intl. Conf. on Collaborative Computing: Networking, Applications and Work sharing, 2011.

[2] C. Dwork, "Differential privacy: a survey of results," in Proc. of the 5th Intl. Conf. on Theory and Applications of Models of Computation, 2008, pp. 1–19.

[3] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Computing Survey., vol. 42, pp. 14:1–14:53, June 2010.

[4] C. Dwork, "A firm foundation for private data analysis," Communication. ACM, vol. 54, pp. 86–95, January 2011.

[5] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," ACM Trans. on Knowledge Discovery from Data, vol. 4, no. 4, pp. 18:1–18:33, October 2010.

[6] W. Jiang and C. Clifton, "Privacy-preserving distributed k-anonymity," in DB Sec, vol. 3654, 2005, pp. 924–924.

[7] W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," VLDB J., vol. 15, no. 4, pp. 316–333, 2006.

[8] O. Goldreich, Foundations of Cryptography: Volume 2, 2004.

[9] Y. Lindell and B. Pinkas, "Secure multiparty computation for privacy-preserving data mining," The Journal of Privacy and Confidentiality, vol. 1, no. 1, pp. 59–98, 2009.

[10] P. Samarati, "Protecting respondents' identities in micro data release," IEEE TKDE, vol. 13, no. 6, pp. 1010–1027, 2001.

[11] L. Sweeney, "k-Anonymity: a model for protecting privacy," Int. J. Uncertain. Fuzziness Knowledge.-Based Syst., vol. 10, no. 5, pp. 557–570,2002.

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy beyond k-anonymity," in ICDE, 2006,p. 24.

[13] R. Burke, B. Mobasher, R. Zabicki, and R. Bhaumik, "Identifying attack models for secure recommendation," in Beyond Personalization: A Workshop on the Next Generation of Recommender Systems, 2005.