A New Profile Based Privacy Measure for Data Publishing

[1]Dr. C.P.V.N.J. Mohan Rao, [2]Kumar Vasantha, [3]HarishBabu. Kalidasu
[1,2,3]Dept. of CSE, Avanthi Institute of Technology & Science, Narsipatnam, Vishakhapatnam, AP, India

Abstract
The k-anonymity privacy requirement for publishing microdata requires that each equivalence class (i.e., a set of records that are indistinguishable from each other with respect to certain "identifying" attributes) contains at least k records. Recently, several authors have recognized that k-anonymity cannot prevent attribute disclosure. The notion of 'diversity has been proposed to address this; l-diversity requires that each equivalence class has at least 'well represented values for each sensitive attribute. In this paper, we follows that l-diversity has a number of limitations. In particular, it is neither necessary nor sufficient to prevent attribute disclosure. Motivated by these limitations, we worked on new notion of privacy called "closeness."In this paper we are introducing performance based automatic data publishing to multiple users using User Profile Category(UPC) , this method enhances the present flexible privacy model called (n,t)-closeness. We discuss the rationale for using closeness as a privacy measure and illustrate its advantages through examples and experiments.

Keywords
Privacy Preservation, Data Anonymization, Data Publishing, Data Security

I. Introduction

GOVERNMENT agencies and other organizations often need to publish microdata, e.g., medical data or census data, for research and other purposes. Typically, such data are stored in a table, and each record (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 Microdata, 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 [8, 15]: 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 [15]. 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. 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. To effectively limit disclosure, we need to measure the disclosure risk of an anonymized table.

While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. To address this limitation of k-anonymity, Machanavajjhala et al. [23] recently introduced a new notion of privacy, called '- diversity, which requires that the distribution of a sensitive attribute in each equivalence class has at least ' "well represented" values.

In this paper, we focusing on a novel privacy notion called "closeness." [a].And we are introducing new method of data publishing .This process enhances the ability of the closeness.

II. Related Work

The problem of information disclosure has been studied extensively in the framework of statistical databases. A number of information disclosure limitation techniques have been designed for data publishing, including Sampling, Cell Suppression, Rounding, and Data Swapping and Perturbation. These techniques, however, insert noise to the data Samarati [30] and Sweeney [32] introduced the k-anonymity model. Since then, there has been a large amount of research work on this topic. We classify them into two categories:

- Privacy measurements
- Anonymization techniques.

A. From k-Anonymity to 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 grater than 1/k.

While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. This has been recognized by several authors, e.g., [23, 33, 40]. Two attacks were identified in [23]: the homogeneity attack and the background knowledge attack. example. To address these limitations of k-anonymity, Machanavajjhala et al.[23] introduced 'l-diversity as a stronger notion of privacy.

Table 1: Original Patients Table

	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 2: A 3-Anonymous Version of Table 1

	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

B. 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. l-diversity may be difficult to achieve and may not provide sufficient privacy protection. l-diversity is insufficient to prevent attribute disclosure. Below, we present two attacks on 'l-diversity. Skewness attack[a], Similarity attack[a].

C. A New Privacy Measure: (n,t)-Closeness

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 population in the released data and about specific individuals.

t-Closeness: Base Model: To motivate our approach, let us perform the following thought experiment[a]. Anonymization Algorithms[a]. One challenge is designing algorithms for anonymizing the data to achieve (n, t)-closeness.

D. Distance Measures

The problem is to measure the distance between two probabilistic distributions. There are a number of ways to define the distance between them[a]

1. Desiderata for Designing the Distance Measure
2. Distance Measure Based on Kernel Smoothing
3. Earth Mover's Distance

III. Proposed Model

In the proposed method, we follow the (n,t)-Closeness, for the privacy preserving and extending the data publishing method while preserving privacy preserving.

A. Data Publishing

Database publishing is an area of automated media production in which specialized techniques are used to generate paginated documents from source data residing in traditional databases. Common examples are catalogues, direct marketing, report generation, price lists and telephone directories. The database content can be in the form of text and pictures but can also contain metadata related to formatting and special rules that may apply to the document generation process. Database publishing can be incorporated into larger workflows as a component, where documents are created, approved, revised and released.

In the proposed method we extending the new privacy measure for data publishing, the system automatically publish the data according to the system requirement and the user profile categories. The data are published dynamically are updated by running predefined macros.

User Profiles are used in conjunction with publications to personalize the content that users see when documents are published using single-pass report bursting, Using profiles; you can schedule a publication, once, and deliver many different personalized versions of the report to users.

Each publishing option has several features:

1. Specifying the data that users see,
2. Allowing users to update the data

For example, you could use a profile to associate regional class information with users and groups, or you could combine the regional information with a profile that provides details about the user's status within the company. To use a profile with Publishing, you need to decide what level of personalization you need and then create the profile and assign it to users and groups. When you schedule and distribute personalized documents through publications, the profile will control what information users see. Profiles do not control users' access to data. Profiles are used to refine a document's content, or filter it. When you use profiles to display a subset of the data to a user, it is not the same as restricting the user from seeing that data. If users have the appropriate rights, they can still see the complete data for the document by viewing the instance.

B. Creating Profiles

1. Personalizing data with profile targets
2. Personalizing data for users and groups

IV. Experiments

The main goal of the experiment is to publish the data dynamically to the servers and to the user profiles. The system incorporates the previous system work nature, The k-anonymity privacy requirement for publishing microdata by using (n,t)-CLOSENESS and using distance measures. In the experiments, we compare four privacy measures with the proposed We compare these privacy measures through an evaluation of 1) vulnerability to similarity attacks; 2) efficiency; and 3) data utility. For each privacy measure, we adapt the Mondrian multidimensional k-anonymity algorithm [17] for generating the anonymized table that satisfies the privacy measure. The data set used in the experiments is the ADULT data set from the UC Irvine machine learning repository [34], which is composed of data collected from the US census? We used seven attributes of the data set, as shown in fig. 3. Six of the seven attributes are treated as quasi-identifiers and the sensitive attribute is Occupation. Records with missing values are eliminated and there are 30,162 valid records in total. The algorithms are implemented in Java and the experiments are run on a 3.4-GHZ Pentium 4 machine with 2 GB memory.

	Attribute	Type	# of values	Height
1	Age	Numeric	74	5
2	Workclass	Categorical	8	3
3	Education	Categorical	16	4
4	Marital_Status	Categorical	7	3
5	Race	Categorical	5	3
6	Gender	Categorical	2	2
7	Occupation	Sensitive	14	3

Fig. 2: Description of the Adult Data Set Used in the Experiment.

V. Conclusions and Futrue Work

While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. The notion of ‘-diversity attempts to solve this problem. We have shown that ‘-diversity has a number of limitations and especially presented two attacks on ‘- diversity. Motivated by these limitations, we have proposed a novel privacy notion called “closeness.” We propose two instantiations: a base model called t-closeness and a more flexible privacy model called (n,t)-closeness. To incorporate semantic distance, we choose to use the Earth Mover Distance measure. We also point out the limitations of EMD, present the desiderata for designing the distance measure, and propose a new distance measure that meets all the requirements. Finally, through experiments on real data, we show that similarity attacks are a real concern and the (n; t)-closeness model better protects the data while improving the utility of the released data. We worked on processing the data publishing.

References

[1] C. Aggarwal,“On k-Anonymity and the Curse of Dimensionality”, Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 901-909, 2005.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D.Thomas, and A. Zhu, “Achieving Anonymity via Clustering”, Proc. ACM Symp. Principles of Database Systems (PODS), pp. 153-162, 2006.

[3] R.K. Ahuja, T.L. Magnanti, J.B. Orlin,"Network Flows: Theory", Algorithms, and Applications. Prentice-Hall, Inc., 1993.

[4] R.J. Bayardo, R. Agrawal,“Data Privacy through Optimal k- Anonymization”, Proc. Int’l Conf. Data Eng. (ICDE), pp. 217-228,2005.

[5] F. Bacchus, A. Grove, J.Y. Halpern, D. Koller,“From Statistics to Beliefs”, Proc. Nat’l Conf. Artificial Intelligence (AAAI), pp. 602- 608, 1992.

[6] J.-W. Byun, Y. Sohn, E. Bertino, N. Li,“Secure Anonymization for Incremental Datasets”, Proc. VLDB Workshop Secure Data Management (SDM), pp. 48-63, 2006.

[7] B.-C. Chen, K. LeFevre, R. Ramakrishnan,“Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge”, Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.

[8] G.T. Duncan, D. Lambert,“Disclosure-Limited Data Dissemination,”J. Am. Statistical Assoc., Vol. 81, pp. 10-28, 1986.

[9] B.C.M. Fung, K. Wang, P.S. Yu,“Top-Down Specialization for Information and Privacy Preservation”, Proc. Int’l Conf. Data Eng. (ICDE), pp. 205-216, 2005.

[10] C.R. Givens, R.M. Shortt,“A Class of Wasserstein Metrics for Probability Distributions”, Michigan Math J., Vol. 31, pp. 231-240, 1984.

[11] V.S. Iyengar,“Transforming Data to Satisfy Privacy Constraints”, Proc. ACM SIGKDD, pp. 279-288, 2002.

[12] D. Kifer, J. Gehrke,“Injecting Utility into Anonymized Datasets”, Proc. ACM SIGMOD, pp. 217-228, 2006.

[13] N. Koudas, D. Srivastava, T. Yu, Q. Zhang, “Aggregate Query Answering on Anonymized Tables”, Proc. Int’l Conf. Data Eng. (ICDE), pp. 116-125, 2007.

[14] S.L. Kullback, R.A. Leibler,“On Information and Sufficiency”, Annals of Math. Statistics, Vol. 22, pp. 79-86, 1951.

[15] D. Lambert,“Measures of Disclosure Risk and Harm”, J. Official Statistics, Vol. 9, pp. 313-331, 1993.

[16] K. LeFevre, D. DeWitt, R. Ramakrishnan,“Incognito: Efficient Full-Domain k-Anonymity”, Proc. ACM SIGMOD, pp. 49-60, 2005.

[17] K. LeFevre, D. DeWitt, R. Ramakrishnan,“Mondrian Multidimensional k-Anonymity”, Proc. Int’l Conf. Data Eng. (ICDE), pp. 25, 2006.

[18] K. LeFevre, D. DeWitt, R. Ramakrishnan, “Workload-Aware Anonymization,” Proc. ACM SIGKDD, pp. 277-286, 2006.

[19] N. Li, T. Li, S. Venkatasubramanian,“t-Closeness: Privacy beyond k-Anonymity and ‘-Diversity”, Proc. Int’l Conf. Data Eng. (ICDE), pp. 106-115, 2007.

[20] T. Li, N. Li,“Injector: Mining Background Knowledge for Data Anonymization”, Proc. Int’l Conf. Data Eng. (ICDE), 2008.

[21] T. Li, N. Li,“Towards OptimalkAnonymization”, Data and Knowledge Eng., Vol. 65, pp. 22-39, 2008.

[22] T. Li, N. Li, J. Zhang,“Modeling and Integrating Background Knowledge in Data Anonymization”, Proc. Int’l Conf. Data Eng.(ICDE), 2009.

[23] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam,"Diversity: Privacy Beyond k-Anonymity”, Proc. Int’l Conf. Data Eng. (ICDE), p. 24, 2006.

[24] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, J.Y. Halpern,“Worst-Case Background Knowledge for Privacy-Preserving Data Publishing”, Proc. Int’l Conf. Data Eng. (ICDE), pp. 126-135, 2007.

[25] A. Meyerson, R. Williams,“On the Complexity of Optimal k- Anonymity”, Proc. ACM Symp. Principles of Database Systems (PODS), pp. 223-228, 2004.

[26] M.E. Nergiz, M. Atzori, C. Clifton,“Hiding the Presence of Individuals from Shared Databases”, Proc. ACM SIGMOD, pp. 665-676, 2007.

[27] H. Park, K. Shim,“Approximate Algorithms for k-Anonymity”, Proc. ACM SIGMOD, pp. 67-78, 2007.

[28] V. Rastogi, S. Hong, D. Suciu,“The Boundary between Privacy and Utility in Data Publishing”, Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 531-542, 2007.

[29] Y. Rubner, C. Tomasi, L.J. Guibas,“The Earth Mover’s Distance as a Metric for Image Retrieval”, Int’l J. Computer Vision, Vol. 40, No. 2, pp. 99-121, 2000.

[30] P. Samarati,“Protecting Respondent’s Privacy in Microdata Release”, IEEE Trans. Knowledge and Data Eng., Vol. 13, No. 6, pp. 1010-1027, Nov./Dec. 2001.

[31] L. Sweeney,“Achieving k-Anonymity Privacy Protection Using Generalization and Suppression”, Int’l J. Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No. 6, pp. 571-588, 2002.

[32] L. Sweeney,“k-Anonymity: A Model for Protecting Privacy”, Int’l J. Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No. 5, pp. 557-570, 2002.

[33] T.M. Truta, B. Vinay,“Privacy Protection: P-Sensitive k-Anonymity Property”, Proc. Int’l Workshop Privacy Data Management (ICDE Workshops), 2006.

[34] A. Asuncion, D.J. Newman,"UCI Machine Learning Repository, [Online] Available: http://www.ics.uci.edu/~mlearn/ML-Repository.html, 2007.

[35] M.P. Wand, M.C. Jones, Kernel Smoothing (Monographs on Statistics and Applied Probability)", Chapman & Hall, 1995.

[36] K. Wang, B.C.M. Fung, P.S. Yu,"Template-Based Privacy Preservation in Classification Problems", Proc. Int'l Conf. Data Mining (ICDM), pp. 466-473, 2005.
[37] R.C.-W. Wong, A.W.-C. Fu, K. Wang, J. Pei,"Minimality Attack in Privacy Preserving Data Publishing", Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 543-554, 2007.
[38] R.C.-W. Wong, J. Li, A.W.-C. Fu, K. Wang,"(_,k)-Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing", Proc. ACM SIGKDD, pp. 754-759, 2006.
[39] X. Xiao, Y. Tao,"Anatomy: Simple and Effective Privacy Preservation", Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.
[40] X. Xiao, Y. Tao,"Personalized Privacy Preservation", Proc. ACM SIGMOD, pp. 229-240, 2006.
[41] X. Xiao, Y. Tao,"m-Invariance: Towards Privacy Preserving Republication of Dynamic Datasets", Proc. ACM SIGMOD, pp. 689-700, 2007.
[42] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A.W.-C. Fu, "Utility-Based Anonymization Using Local Recoding", Proc. ACM SIGKDD, pp. 785-790, 2006.