Literature Survey for Secure Anonymization

Ishwarya M.V, K.Ramesh Kumar

Abstract- In this paper we are going to discuss about the privacy preservation of Patients details in a medical centre . The medical centre may have various login for various people like Administrator,Doctor,Analyst and Receptionist .We design a model such that the patients entire details are not known to everyone who logins with their id.It is available in a suppressed form to each and everyone who logs in .

All the patients data are being split using Slicing algorithm and shuffled and stored in different databases in encryption side.The data are realigned and deshuffled and the original data are retrieved in the decryption side.

Keywords: Privacy preservation, Authentication, Security, Slicing, Shuffling.

LITERATURE SURVEY

1. R.J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization," Proc. Int'l Conf. Data Eng. (ICDE), pp. 217-228, 2005.

Data de-identification reconciles the demand for release of data for research purposes and the demand for privacy from individuals. This paper proposes and evaluates an optimization algorithm for the powerful de-identification procedure known as k-anonymization. A k-anonymized dataset has the property that each record is indistinguishable from at least others. Even simple restrictions of optimized - anonymity are NP-hard, leading to significant computational challenges. We present a new approach to exploring the space of possible anonymizations that tames the combinatorics of the problem, and develop data-man-agement strategies to reduce reliance on expensive opera-tions such as sorting. Through experiments on real census data, we show the resulting algorithm can find optimal k-anonymizations under two representative cost measures and a wide range of . We also show that the algorithm can produce good anonymizations in circumstances where the input data or input parameters preclude finding an optimal solution in reasonable time. Finally, we use the algorithm to explore the effects of different coding approaches and problem variations on anonymization quality and performance. To our knowledge, this is the first result demonstrating optimal -anonymization of a non-trivial dataset under a general model of the problem.

2. F. Bacchus, A. Grove, J.Y. Halpern, and D. Koller, "From Statistics to Beliefs," Proc. Nat'l Conf. Artificial Intelligence (AAAI), pp. 602-608, 1992.

An intelligent agent uses known facts, including statistical Knowledge, to assign degrees of belief to assertions it is uncertain about. We investigate three principled techniques for doing this. All three are applications of the principle of Indifference, because they assign equal degree of belief to all basic "situations" consistent with the knowledge base. They differ because there are competing intuitions about what the basic situations are. Various natural patterns of reasoning, such as the preference for the most specific statistical data available turn out to follow from some or all of the techniques. This is an improvement over earlier theories, such as work on direct inference and reference classes, which arbitrarily postulate these patterns without offering any deeper explanations or guarantees of consistency. The three methods we investigate have surprising characterizations there are connections to the principle of maximum entropy, a principle of maximal independence, and a "center of mass" principle. There are also unexpected connections between the three that help us understand why the specific language chosen (for the knowledge base) is much more critical in inductive reasoning of the sort we consider than it is in traditional deductive reasoning.

3. J.-W. Byun, Y. Sohn, E. Bertino, and N. Li, "Secure Anonymization for Incremental Datasets," Proc. VLDB Workshop Secure Data Management (SDM), pp. 48-63, 2006.

Data anonymization techniques based on the k-anonymity model have been the focus of intense research in the last few years. Although the k-anonymity model and the related techniques provide valuable solutions to data privacy, current solutions are limited only to the static data release (i.e., the entire dataset is assumed to be available at the time of release). While this may be acceptable in some applications, today we see databases continuously growing everyday and even every hour. In such dynamic environments, the current techniques may suffer from poor data quality and/or vulnerability to inference. In this paper, we analyze various inference channels that may exist in multiple anonymized datasets and discuss how to avoid such inferences. We then present an approach to securely anonymizing a continuously growing dataset in an efficient manner while assuring high data quality.

4. N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-Anonymity and '-Diversity," Proc. Int'l Conf. Data Eng. (ICDE), pp. 106-115, 2007.

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 l-diversity has been proposed to address this; l-diversity requires that each equivalence class has at least l well-represented values for each sensitive attribute. In this paper we show that l-diversity has a number oflimitations. In particular, it is neither necessary nor sufficient to prevent attribute disclosure. We propose a novel privacy notion called t-closeness, which requires 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). We choose to use the Earth Mover Distance measure for our t-closeness requirement. We

Manuscript received December, 2013

Ishwarya M.V, Research Scholar,HITS,Padur,Assistant Professor,Sri Sairam Engg.College,

Dr K.Ramesh Kumar, Associate Professor,HITS,Padur

discuss the rationale for t-closeness and illustrate its advantages through examples and experiments.

5. T. Li and N. Li, "Injector: Mining Background Knowledge for Data Anonymization," Proc. Int'l Conf. Data Eng. (ICDE), 2008.

Existing work on privacy-preserving data publishing cannot satisfactorily prevent an adversary with background knowledge from learning important sensitive information. The main challenge lies in modeling the adversary's background knowledge. We propose a novel approach to deal with such attacks. In this approach, one first mines knowledge from the data to be released and then uses the mining results as the backgroundknowledge when anonymizing the data. The rationale of our approach is that if certain facts or background knowledge exist, they should manifest themselves in the data and we should be able to find them using data mining techniques. One intriguing aspect of our approach is that one can argue that it improves both privacy and utility at the same time, as it both protects against background knowledge attacks and better preserves the features in the data. We then present the Injector framework for data anonymization. Injector mines negative association rules from the data to be released and uses them in the anonymization process. We also develop an efficient anonymization algorithm to compute the injected tables that incorporates background knowledge. Experimental results show that Injector reduces privacy risks against background knowledge attacks while improving data utility.

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, and J.B. Orlin, Network Flows: Theory, Algorithms, and Applications. Prentice-Hall, Inc., 1993.

[4] R.J. Bayardo and 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, and 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, and N. Li, "Secure Anonymization for Incremental Datasets," Proc. VLDB Workshop Secure Data Management (SDM), pp. 48-63, 2006.

[7] B.-C. Chen, K. LeFevre, and 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 and D. Lambert, "Disclosure-Limited Data Dissemination," J. Am. Statistical Assoc., vol. 81, pp. 10-28, 1986. LI ET AL.: CLOSENESS: A NEW PRIVACY MEASURE FOR DATA PUBLISHING 955

[9] B.C.M. Fung, K. Wang, and 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 and 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 and J. Gehrke, "Injecting Utility into Anonymized Datasets," Proc. ACM SIGMOD, pp. 217-228, 2006.

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

[14] S.L. Kullback and 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.