ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 3, Issue 5, May 2015

Privacy Preserving Issues in Data Publishing

Kavitha S[1], Priyadharshini V[2], Yamini S[3]

PG Scholars, Dr. N.G.P. Institute of Technology, Coimbatore, India[1,2,3]

Abstract: Data publishing is recently focused more for the data analysis. In recent days, the data creation is enormous in all the fields. Predictive analytics uses all the data collected from various sources for predicting the future even though there is uncertainty in information gathered. The data used for analysis should not affect the privacy for the record owners. In all the sectors, organizations use their data for predictive analytics. The organizations should not reveal the sensitive details of the record owners for any cause. In general the data privacy is preserved with data anonymization. There are algorithms such as k-anonymity, l-diversity and t-closeness for data anonymization. On anonymizing the data, there are threats that can disclose the information about the record owner. In this paper we discussed about the threats that affects the privacy of record owners in data publishing.

Keywords: privacy, data publishing, privacy threats, anonymization, privacy protection.

I. INTRODUCTION

The data publishing is idea for the effectiveness of data utility. Privacy preserving data publishing [1] ensures the data privacy without compromising the data utility. The data utility should be remained as it is even after anonymizing the data. Techniques like Slicing [2] also help to maintain the data utility with the data privacy. Data anonymization provides generalization and suppression with k-anonymity, t-closeness and l-diversity algorithms [3], [4], [5]. Generalization replaces the data with less specific values that can be used for analysis whereas Suppression changes the complete value that cannot be used for any other purposes [6], [7], [8].

For example, the attribute Name can be suppressed because it encloses the identity of a person. Perturbation techniques are also used for data privacy. These techniques add noise to the data so that the data can be retrieved by other person only if the added noise is known. But Perturbation techniques cannot be used for data publishing because it is similar to encryption which deals with data confidentiality [9]. These techniques also do not provide complete privacy. In this paper we discussed about the threats in privacy preserving data publishing.

II. THREATS TO PPDP

The data anonymization and other techniques are used for privacy preserving data publishing but the anonymized data also have the threats that can disclose the individual’s information. The data anonymization mainly involves attribute and membership disclosure [10]. In web search there is a chance of identity disclosure which are protected by personalized web search [11], [12]. The data recipient receiving data from the publisher could be an attacker. The challenge in Privacy Preserving Data Publishing is to maintain the privacy and also to perform data anonymization. The following threats are the challenges in privacy preserving data publishing.

III. IDENTITY DISCLOSURE

Identity disclosure is the identification of an individual with the data values in the data set by other users.

A. Khaled El Emam and Fida Kamal Dankar [13]

The data anonymization using k-anonymity can disclose individual information (identity disclosure). The risk on re identification of an individual in k-anonymity is evaluated in this paper. In k-anonymity k should be at least less than k-1 records for identifying the variables. When k=10 with the values of potentially identifying variables then at least 10 records would be of same combination that can disclose the information of an individual. The threshold risk is when the probability of k value is maximum of being 1/k. It results in re-identification of an individual. In this paper, the re-identification of an individual is evaluated with threshold risk provides more information loss and the re-identification of an arbitrary individual produces low information loss.

B. Dan Zhu, Xiao-Bai Li and Shuning Wu [14]

The most serious privacy concern in data publishing is identity disclosure. The familiar algorithm for identity disclosure protection is k-anonymity. The drawback of this algorithm is the third party can reveal the individual information from the anonymized data set. In this paper the privacy of identity disclosure is provided with data reconstruction technique that helps the privacy preserved anonymized data to be used for predictive analytics. The data reconstruction is done with aggregation and swapping of the data. The aggregation is done for numerical data whereas swapping is done for nominal data. From the reconstructed data, the subset of data is identified with genetic algorithm and further replication of subset is done for satisfying the constraints of k-anonymity. This results in more protection over identity disclosure issue in data anonymization.

C. Kun Liu and Evimaria Terzi [15]

The data perturbation and anonymization techniques for tabular data are easier than with graphs. In the graph notation, edges and vertices are considered and with this re-identification of an individual is protected with the background knowledge. The anonymization is done with minimum number of nodes and edges. In this paper the problem is split into two sub problems, one is to anonymize the data and the other is addition and deletion

DOI 10.17148/IJIREEICE.2015.3504

of the nodes and edges in the graph. The addition and deletion of even a single node or edge affects the whole network which should be taken care of and it is tedious process. In this method the privacy is enhanced but the data utility is difficult to measure.

IV. MEMBERSHIP DISCLOSURE

Membership disclosure is the identification of one particular person’s data is added in the published data set by the adversaries

D. Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu and Weining Yang [16]

The membership disclosure is not protected by l-diversity and t-closeness. Even though data anonymization provides protection for membership disclosure it is not completely protected. In this paper to enhance the privacy for membership disclosure a membership privacy framework is developed. This new framework consists of positive and negative membership privacy. The ability of finding the particular data in the published data by the adversaries is protected in positive membership privacy and the leakage of non-membership privacy is protected with negative membership privacy. This is implemented using the background knowledge of the adversaries. Differential privacy and differential identity is also used for membership disclosure protection under sampling. These are used in this framework for better understanding of the relationships between the data and also for better data utility.

V. ATTRIBUTE DISCLOSURE

Attribute disclosure is the identification of sensitive information of an individual from the set of aggregated data.

A. Justin Brickell and Vitaly Shmatikov [17]

In this paper the threats to the data privacy and utility is discussed. The data privacy produced by the generalization and suppression algorithms like k-anonymity, l-diversity, t-closeness affects the data utility. The privacy risk is determined from the attributes(quasi-identifiers) linked with the sensitive values called attribute disclosure. This is different from membership disclosure. If the quasi identifiers and sensitive values are not published together, then it can be prevented. Here the experiment results on applying semantic privacy to the data for generalization and suppression considerably affects the data utility. The other algorithms k-anonymity and l-diversity does not protect the data from attribute disclosure. The complete protection from t-closeness is not provided but it in directly it provides less privacy from attribute disclosure.

B. Xiaoxun Sun, Lili Sun and Hua Wang [18]

The top down specification algorithm is used for generalization to preserve privacy while data publishing. The top down specification is a k-anonymity algorithm which generalizes the data from parent node to the child node. The k-anonymity does not prevent attribute disclosure. This is enhanced with P-Sensitive k-anonymity model. Even this model does not provide the enough privacy over attribute disclosure. In this paper, a new model called p^+ Sensitive k-anonymity is proposed to enhance the privacy over attribute disclosure. The efficiency and effectiveness of this model with top down specification is measured with ordinal metric system. The sensitive values with QI attributes are not disclosed together and also how much QI attributes contributes to the sensitive value measures the attribute disclosure threat in privacy preserving data publishing. The proposed method is comparatively better than p-Sensitive k-anonymity and k-anonymity model in preserving attribute disclosure.

C. Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu and Ke Wang [19]

The identity disclosure and the attribute disclosure are protected by (α,k) anonymity model. In this model, when two or more sensitive value are there in the data set then it is combined into one class and the (α,k) anonymity model is applied. From the experiment of real world data it is inferred that the enhance model provides better privacy than general (α,k) anonymity. The (α,k) anonymity can be used for the data set in which all the attributes are sensitive and needs to be protected. It is used when all sensitive attribute has many values and not any single value dominates the attribute. The distribution of the attribute should be even for the inferring the confidence in equivalent class.

VI. CONCLUSION

The growth of data publishing leads to more privacy concerns about the data. The data provided by an individual to an organization should be prevented from all types of security issues. The data that is stored without any purpose is not meaningful. All the data generated in the world is used for predictions. These data should be privacy preserved before publishing and used for other purposes. The privacy preservation is implemented with various methods but still there are threats in preserving privacy. In this paper, we discussed about the privacy issues like identity disclosure, membership disclosure, attribute disclosure and other alternate ways to overcome the privacy threats in data publishing.

These threats are not only in data publishing but also threaten web services especially in web search. The personalized web search also gets affected with identity disclosure when the user profile is created but it is not discussed in this paper.

REFERENCES

[1]. 1. B. C. M. FUNG, KE WANG, RUI CHEN, PHILIP S. YU, "Privacy-Preserving Data Publishing: A Survey of Recent Developments", ACM Computing Surveys, Vol. 42, No. 4, June 2010.

[2]. .Tiancheng Li, Ninghui Li, Jian Zhang and Ian Molloy (2012), “Slicing: A New Approach for Privacy Preserving Data Publishing”, IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 3.

[3]. 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.

DOI 10.17148/IJIREEICE.2015.3504

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

[5]. N. Li, T. Li, and S. Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity", in Proc. of the IEEE ICDE (2007), pp. 106–115.

[6]. Latanya Sweeney, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, "Achieving k-anonymity privacy protection using generalization and suppression", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Volume 10 Issue 5, October 2002, Pages 571-588.

[7]. P. Samarati, "Protecting respondants' identities in microdata release", IEEE Transactions on Knowledge and Data Engineering, 13(6), November/December 2001.

[8]. P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.

[9]. Stanley, R. M., Oliveira and Osmar R. Za ¨iane, "Privacy Preserving Clustering by Data Transformation", Journal of Information and Management", Vol. 1, No. 1, 2010.

[10]. X. Xiao and Y. Tao, 'Personalized Privacy Preservation', Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.

[11]. Lidan Shou, He Bai, Ke Chen, and Gang Chen, 'Supporting Privacy Protection in Personalized Web Search,' IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 2, February 2014.

[12]. X. Shen, B. Tan, and C. Zhai, 'Privacy Protection in Personalized Search', SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.

[13]. Khaled El Emam and Fida Kamal Dankar, "Protecting Privacy Using k-Anonymity",J Am Med Inform Assoc. 2008 Sep-Oct; 15(5): 627–637.

[14]. Dan Zhu, Xiao-Bai Li and Shuning Wu,"Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining",J. Decision Support Systems Volume 48, Issue 1, December 2009, Pages 133–140.

[15]. Kun Liu and Evimaria Terzi, "Towards Identity Anonymization on Graphs",SIGMOD'08,June 9–12, 2008, Vancouver, BC, Canada.

[16]. Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu and Weining Yang, "Membership privacy: a unifying framework for privacy definitions",Proceeding CCS '13 Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security Pages 889-900.

[17]. Justin Brickell and Vitaly Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing",KDD'08,August 24–27, 2008.

[18]. Xiaoxun Sun, Lili Sun and Hua Wang, Extended k-anonymity models against sensitive attribute disclosure, J. Computer Communications Volume 34, Issue 4, 1 April 2011, Pages 526-535.

[19]. Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu and Ke Wang ,"(a, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing",KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

DOI 10.17148/IJIREEICE.2015.3504