ISSN No:-2456-2165

Crowd-Sourced Data Publishing Using Untrusted Server

Laya Chacko[1]
Computer Science and Engineering
Mount Zion College of Engineering Pathanamthitta

Bibin Varghese[2]
Computer Science and Engineering
Mount Zion College of Engineering, Pathanamthitta

Smita C Thomas[3]
Computer Science and Engineering
Mount Zion College of Engineering, Pathanamthittta

Abstract:- This paper outlines statistical publishing of real time data with strong protection. A person has sensitive, non-sensitive data which needs protection when publishing it to public. For example health data, it includes disease nature, health condition, medicine prescription, patient name and details these are sensitive data which may be leaked when publishing. In this, certain techniques are applied when transmitting to public here introduces distributors and agents. Propose a privacy preserving framework called privacy preserving distributed agent P2DA. Various techniques are included for prevent information loss they are level-based, multi-set generalization, K-anonymity.

Keywords:- *Differential Privacy, Level-Based, Multi-Set Generalization, K-Anonymity, Data Publishing.*

I. INTRODUCTION

Explosive growth in volume and variety of data's generated due to the various applications. Various agencies such as government and several organisation publish data for research and data mining purposes. These data's are stored in table as rows and columns. Crowd sourced data from millions of users used for discover valuable information. However, data publishing is a risky factor that linked back to the individuals. In digitalised world, privacy of individuals is a challenging factor. To protect against various privacy threats introduce several privacy preservation techniques.

The existing privacy preservation of data publishing are ε-differential privacy and ω- event privacy. ω-event privacy, which protects event sequence in successive time stamps and ε-privacy provide one time statistical data publishing. Traditional system follows trusted server for data collection and publishing. These trusted server maybe hacked and becomes unsecure. Then the identity and sensitive information be exposed.

During this paper, focus on crowd-sourced real time data publishing. A distributed agent based privacy preservation for data publishing using untrusted server. In this multiple agents are introduced between the users and untrusted server. A users can randomly select one representative for transmit the data to it using broadcast mechanism. Along with ω-event privacy and ε-differential privacy several other techniques are introduced level based, multi-set generalization, K-anonymity methods for privacy preservation.

II. RELATED WORK

Protection on statistical data publishing proposed several techniques. Hong et al. [3] proposed K-anonymity in publishing users logs. Winslett et.al. [1] realized one-time statistical data release for differentially private data publishing. Montjoye et al. [4], proposed differentially private spatial decomposition techniques to partition the space. Social network data publishing disclosure problem address using randomized perturb method. Ganti et al. [5] proposed privacy preserving time series data for statistics publishing. Xiong et.al. [7], proposed a real time monitoring of differential privacy using adaptive approach.

Chen et al. [6] proposed a participant-density-aware privacy-preserving aggregate statistics scheme, making use of multi-pseudonym mechanism. X. Lu [2], proposed a spatial-temporal crowd-sourced data publishing in real time with differential privacy.

III. PROPOSED SYSTEM

The volcanic growth of data from various applications are collected by various data collectors. The finest example is healthcare organization. The healthcare organization can use statistical analysis to extract valuable information about its patients for research purposes and doctor's has able to retrieve patient's medical records. Different organizations done the research, and the control over the individual data is hard to enforce. Also it will be used in a wrong way.

The proposed system consist of two section: the agent and the distributor. Here, the distributor registers the agent and also determines priority of each agent during the time of registration. Each of the individual has its own priority that assigned at the time of registration and distributor access request from registered agents. The agent is in between user and the distributor. The agent who has direct connection with the untrusted server and the requested

database by the agent is specify. Next, tabulate the sensitive and non-sensitive attributes. The related rule set is applied based on the level assigned by the distributor to each agent. The rule sets are depict on the basis of the degree of data hidden to various type of agent. Hence the transformed dataset relay using routing mechanism and will reach to the agent who requests the dataset.

Datas stored in the database table and the row corresponds to one individual. Attributes in this table classified into: sensitive attributes, attributes easily identify individual and the values taken together identify an individual. Different agencies and organisation release data for various data mining purpose.

Anonymization techniques like generalisation, bucketization, and slicing. The first and foremost step is attribute partitioning and next is removing of explicit identifiers. Generalisation which replaces quasi-identifiers values with less specific values. However, k-anonymity protects the identification disclosure and generalisation has considerable amount of information loss. A new level of privacy was introduced for address homogeneity attack called level based-diversity (l-diversity) there must be "l" well represented values for the sensitive attributes. Records are allotted based on the count of sensitive attributes happening and group the similar records and analyse it. After check the diversity integrate the set of correlated attributes. Bucketization doesn't hinder level based diversity and it is possible to gain information about sensitive attribute as long as information about global distribution.

In slicing, to protect membership disclosure we use column generalization and extremely fit attributes are placed in the same column after attribute partitioning. Two type's data structures are used for tuple partitioning

- Buckets queue
- Set of sliced buckets

At first, only one bucket in the queue which contains tuples and the sliced buckets are empty. For each execution of algorithm, buckets are removed from the queue and splits into two buckets. The two buckets are placed at the end of queue if sliced table satisfy l-diversity. Else it does not split bucket and place in the sliced bucket. The sliced table is calculated when queue empty.

Slicing has the ability to handle high dimensional data by splitting into columns and slicing reduces data dimensionality. Each column viewed as sub-table with lower dimensionality and these sub-tables are linked by the buckets in slicing.

IV. CONCLUSION

The proposed software can be placed in any firm and can hold any databases. This proposed software focus on security and privacy preservation. Different privacy preservation rulesets are apply based on priority level. Compared to the existing system, introduce multiple agents in between user and server. i.e, agent based privacy framework for avoiding privacy leakage. The implementation of level based privacy method, multiset method and K-anonymity reduces the information loss issue.

REFERENCES

[1]. J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett, "Differentially private histogram publication," The VLDB Journal, vol. 22, no. 6, pp. 797– 822, 2013.
[2]. Q. Wang, Y. Zhang, X. Lu, Z. Wang, Z. Qin, and K. Ren, "Rescuedp: Realtime spatio-temporal crowd-sourced data publishing with differential privacy," in Proc. of IEEE INFOCOM, 2016, pp. 1–9.
[3]. Y. Hong, X. He, J. Vaidya, N. Adam, and V. Atluri, "Effective anonymization of query logs," in Proc. of ACM CIKM. ACM, 2009, pp. 1465–1468.
[4]. G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu, "Differentially private spatial decompositions," in Proc. of IEEE ICDE. IEEE, 2012, pp. 20–31.
[5]. R. K. Ganti, N. Pham, Y.E. Tsai, and T. F. Abdel zaher, "Poolview: stream privacy for grassroots participatory sensing," in Proceedings of ACM SenSys. ACM, 2008, pp. 281–294.
[6]. J.Chen, H.Ma, D.S.Wei, and D.Zhao, "Participant-density-aware privacypreserving aggregate statistics for mobile crowd-sensing," in Proc. of IEEE ICPADS. IEEE, 2015, pp. 140– 147.
[7]. L. Fan and L. Xiong, "An adaptive approach to real-time aggregate monitoring with differential privacy,"IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2094–2106, 2014.