International Conference on "Emerging Trends in Computer Engineering, Science and Information Technology"- 2015
Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477

Design And Analysis of Three Party Protocol to Retrieve Private Data By Using Vertical Partitioned Data

Ms. Gauri V. Sonawane **Prof. Naresh Thoutam**

***Abstract* — Design And Analysis Of Three Party Protocol to Retrieve Private Data By Using Vertical Partitioned Data is for Privacy- Privacy-preserving data publishing addresses the problem of disclosing sensitive data when mining for useful information. Among the existing privacy models, e-differential privacy provides one of the strongest privacy guarantees. In this article,address the problem of private data publishing, where different attributes for the same set of individuals are held by two parties. In particular,present an algorithm for differentially private data release for vertically partitioned data between two parties in the semihonest adversary model. To achieve this, first present a two-party protocol for the exponential mechanism. This protocol can be used as a subprotocol by any other algorithm that requires the exponential mechanism in a distributed setting. Furthermore, we propose a three-party algorithm that releases differentially private data in a secure way according to the definition of secure multiparty computation. Experimental results on real-life data suggest that the proposed algorithm can effectively preserve information for a data mining task.of this approach is given in standard model. Public-key patient-controlled encryption structure yet to be known.**

***Key Words* — Differential privacy, secure data integration, classification analysis**

I. INTRODUCTION

Huge databases exist today due to the rapid advances in communication and storing systems. Each database is owned by a particular autonomous entity, for example, medical data by hospitals, income data by tax agencies, financial data by banks, and census data by statistical agencies. Moreover, the emergence of new paradigms such as cloud computing increases the amount of data distributed between multiple entities. These distributed data can be integrated to enable better data analysis for making better decisions and providing high-quality services. For example, data can be integrated to improve medical research, customer service, or homeland security. However, data integration between autonomous entities should be conducted in such a way that no more information than necessary is revealed between the participating entities. At the same time, new knowledge that results from the integration process should not be misused by adversaries to reveal sensitive information that was not available before the data integration.The propose an algorithm to securely integrate person-specific sensitive data from three data providers, whereby the integrated data still retain the essential information for supporting data mining tasks. The following real-life scenario further illustrates the need for simultaneous data sharing and privacy preservation

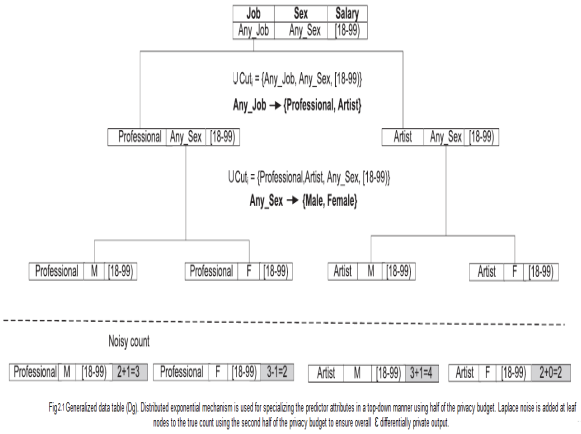


Fig.2 Generalized data table (Dg). Distributed exponential mechanism is used for specializing the predictor attributes in a top-down manner using half of the privacy budget. Laplace noise is added at leaf nodes to the true count using the second half of the privacy budget to ensure overall ε differentially private output.

Fig. 1. Genaralised Data Disrtributed

of person-specific sensitive data. For Example in bank system customer information related to salary transactions may store at one site and information related to loan transactions, Credit Card transaction may store at different database with same customer id. In some cases even different firms like banks and insurance companies may also want share their customer information to study common patterns. This kind of applications falls in to vertically partitioned applications where with same id different set of attributes stored at different sites and common patterns among such sites need to be extracted. A lot of research work is progressing in transforming conventional centralized data mining applications to handle vertically partitioned databases.Also Provide the sensitive information in Encrypted Format.

These technique consist of five Three-Party Algorithm algorithm.There are three functional requirements compactness and correctness or accuracy.Privacy Model Security Model,Data Encryption Algorithm.

II. RELATED WORK

Data privacy has been an active research topic in the statistics, database, and security communities for the last three decades [17]. The proposed methods can be roughly categorized according to two main scenarios:Interactive versus noninteractive. In an interactive framework, a data miner can pose queries through a private mechanism, and a database owner answers these queries in response. In a noninteractive frame-work, a database owner first anonymizes the raw data and then

releases the anonymized version for data analysis. Once the data are published, the data owner has no further control over the published data. This approach is also known as privacy-preserving data publishing (PPDP) [17].Single versus multiparty. Data may be owned by a single party or by multiple parties. In the distributed (multiparty) scenario, data owners want to achieve the same tasks as single parties on their integrated data without sharing their data with others. This proposed algorithm addresses the distributed and noninteractive scenario. Below, we briefly review the most relevant research works.Single-party scenario. We have already discussed different privacy models.Here,we provide an overview of some relevant anonymization algorithms. Many algo-rithms have been proposed to preserve privacy, but only a few have considered the goal for classification analysis [17]. Iyengar [25] has presented the anonymity problem for classification and proposed a genetic algorithmic solution. Bayardo and Agrawal [3] have also addressed the classifi-cation problem using the same classification metric of [25]. Fung et al. [18] have proposed a top-down specialization (TDS) approach to generalize a data table. LeFevre et al have proposed another anonymization technique for classification using multidimensional recoding. More discussion about the partition-based approach can be found in the survey of Fung et al. [17].

Differential privacy [14] has recently received consider-able attention as a substitute for partition-based privacy models for PPDP. However, so far most of the research on differential privacy concentrates on the interactive setting with the goal of reducing the magnitude of the added noise [11], [14], [18], releasing certain data mining results [4], [8], [9], [16], or determining the feasibility and infeasibility results of differentially-private mechanisms [5], [13], [16]. Research proposals [2], [23],that address the problem of noninteractive data release only consider the single-party scenario. Therefore, these techniques do not satisfy the privacy requirement of our data integration application for the financial industry. A general overview of various research works on differential privacy can be found in the survey of Dwork [12]. Distributed interactive approach. This approach is also referred to as privacy preserving distributed data mining (PPDDM) [10]. In PPDDM, multiple data owners want to compute a function based on their inputs without sharing their data with others. This function can be as simple as a count query or as complex as a data mining task such as classification, clustering, and so on. For example, multiple hospitals may want to build a data mining model for predicting disease based on patients' medical history without sharing their data with each other. In recent years, different protocols have been proposed for different data mining tasks including association rule mining, clustering [11], and classification [33], [6]. However, none of these methods provide any privacy guarantee on the computed output (i.e., classifier, association rules). On the other hand, Dwork et al. [13], and Narayan and Haeberlen have proposed interactive algorithms to compute differentially private count queries from both horizontally and vertically partitioned data, respectively.

However, when compared to an interactive approach, a noninteractive approach gives greater flexibility because data recipients can perform their required analysis and data exploration, such as mining patterns in a specific group of records, visualizing the transactions containing a specific pattern, or trying different modeling methods and parameters.

Distributed noninteractive approach. This approach allows anonymizing data from different sources for data release without exposing the sensitive information. Jurczyk and Xiong [27] have proposed an algorithm to securely integrate horizontally partitioned data from multiple data owners without disclosing data from one party to another. Mohammed et al. [41] have proposed a distributed algorithm to integrate horizontally partitioned high-dimensional health care data. Unlike the distributed anonymization problem for vertically partitioned data studied in this paper, these methods [27], [11] propose algorithms for horizontally partitioned data. Jiang and Clifton [26] have proposed the Distributed k-Anonymity (DkA) framework to securely integrate two data tables while satisfying the k-anonymity requirement. Mohammed et al.have proposed an efficient anonymi-zation algorithm to integrate data from multiple data owners. To the best of our knowledge, these are the only two methods [25]that generate an integrated anon-ymous table for vertically partitioned data. However, both methods adopt k-anonymity or its extensions as the underlying privacy principle and, therefore, both are vulnerable to the recently discovered privacy attacks [15], [19]. Table 2 summarizes the different characteristics of the PPDP algorithms discussed above.

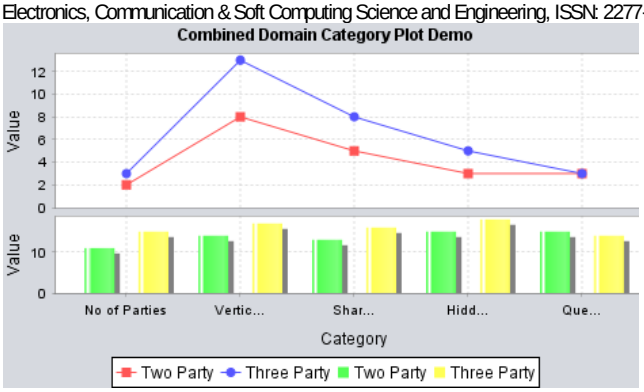
A Formal Statement of Our Contribution Our contribution, as suggested by the paper's title, comes in the parts of privacy, accuracy, and consistency, each of which are critical components of any data analysis system. At an intuitive level, which we soon formalize, we are concerned with

Privacy: The presence or absence of any one data element should not substantially influence the distribution over outcomes of the computation. Accuracy: The difference between the reported marginals and true marginals should be bounded, preferably independent of the size of the data set.Consistency: There should exist a contingency table whose marginals equal the reported marginals. Partition-based approach divides a given data set into disjoint groups and releases some general information about the groups. The two most popular anonymization techniquesare generalization and bucketization. Generalization makes information less precise while preserving the "truthfulness" of information. Unlike generalization,bucketization does not modify the QID and the sensitive attribute (SA) values but instead de-associates the relationship between the two. However, it thus also disguises the correlation between SA and other attributes and, therefore,hinders data analysis that depends on such correlation.Many algorithms have been proposed to preserve privacy,but only a few have considered the goal for classification.Iyengar presents the anonymity problem for classificationand proposes a genetic algorithmic solution. Bayardo and Agrawal also address the classification problem using the same classification metric of. Fung et al. proposea top-down specialization (TDS) approach to generalize a data table. Recently, LeFevre et. al. propose another anonymization technique for classification using multidimensional recoding. All these algorithms adopt k-anonymity or its extensions as the underlying privacy principle and, therefore, are vulnerable to the recently discovered privacy attacks. More discussion about the partition-based approach can be found in a survey paper.Differential privacy has received considerable attention

recently as a substitute for partition-based privacy models for PPDP. However, most of the research on differential privacy so far concentrates on the interactive setting with the goal of reducing the magnitude of added noise,releasing certain data mining results or determining the feasibility and infeasibility results of differentially private mechanisms . A general overview of various research works on differential privacy can be found in the recent survey. Below, we briefly review the results relevant to this paper.Barak et al.address the problem of releasing a set of consistent marginals of a contingency table. Their method ensures that each count of the marginals is non-negative and their sum is consistent for a set of marginals. Xiao et al.propose Privelet, a wavelet-transformation-based approach that lowers the magnitude of noise needed to ensure differential privacy to publish a multidimensional frequencymatrix. Hay et al. propose a method to publish differentially private histograms for a one-dimensional data set.Although Privelet and Hay et al.'s approach can achieve differential privacy by adding polylogarithmic noise variance,the latter is only limited to a one-dimensional data set.Some works address how to compute the results of a number of given queries while minimizing the added noise.However, these methods require the set of queries to be given first altogether to compute the results. In contrast, our method complements the above works by determining how to partition the data adaptively so that the released data can be useful for a given data mining task. In addition, a number of recent works propose differentially-private mechanisms for different applications such as record linkage, and recommender systems. Though closely related, all these works do not address the problem of privacy-preserving data publishing for classification analysis.

III. ANALYSIS

In this evaluate the scaling impact on the data utility in terms of classification accuracy.Then compare DistDiffGen with DiffGen and with the distributed algorithm for k-anonymity ,which we, henceforth, refer to as DAKA. The algorithm DAKA integrates and publishes distributed data with k-anonymity privacy guarantee for classification analysis. Finally,estimate the computation and the communication costs of DistDiffGen. Employ the publicly available data set Adult [15], [18], a real-life census data set that has been used for testing many anonymization algorithms [3], [18], [25]. It has 45,222 census records with six numerical attributes, eight categorical attributes, and a binary class column representing two income levels,50K or >50K. All experiments are conducted on an Intel Core i7 2:7-GHz PC with 12-GB RAM.



CONCLUSION

In this paper, presented the first three-party differentially private data release algorithm for vertically partitioned data. We have shown that the proposed algorithm is differentially private and secure under the security definition of the semihonest adversary model. Moreover, we have experimentally evaluated the data utility for classification analysis. The proposed algorithm can effectively retain essential information for classification analysis. It provides similar data utility compared to the recently proposed single-party algorithm [38] and better data utility than the distributed k-anonymity algorithm for classification analysis [39].

ACKNOWLEDGEMENT

The authors would like to acknowledge Computer Engineering department, SITRC and all the people who provided with the facilities being required and conductive conditions for completion of the review paper.

REFERENCES

[1]R. Agrawal, A. Evfimievski, and R. Srikant, "Information Sharing Across Private Databases," Proc. ACM Int'l Conf. Management of Data, 2003.
[2]B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, "Privacy Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release," Proc. ACM Symp. Principles of Database Systems (PODS '07), 2007.
[3]R.J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization," Proc. IEEE Int'l Conf. Data Eng. (ICDE '05),
2005.
[4]R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, "Discovering Frequent Patterns in Sensitive Data," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '10), 2010.
[5]A. Blum, K. Ligett, and A. Roth, "A Learning Theory Approach to Non-Interactive Database Privacy," Proc. ACM Symp. Theory of Computing (STOC '08), 2008.
[6]J. Brickell and V. Shmatikov, "Privacy-Preserving Classifier Learning," Proc. Int'l Conf. Financial Cryptography and Data Security, 2009.
[7]P. Bunn and R. Ostrovsky, "Secure Two-Party K-Means Cluster-ing," Proc. ACM Conf. Computer and Comm. Security (CCS '07), 2007.
[8]K. Chaudhuri, C. Monteleoni, and A. Sarwate, "Differentially Private Empirical Risk Minimization," J. Machine Learning Re-search, vol. 12, pp. 1069-1109, July 2011.
[9]K. Chaudhuri, A.D. Sarwate, and K. Sinha, "Near-Optimal Differentially Private Principal Components," Proc. Conf. Neural Information Processing Systems, 2012.

[10]C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M.Y. Zhu, "Tools for Privacy Preserving Distributed Data Mining," ACM SIGKDD Explorations Newsletter, vol. 4, no. 2, pp. 28-34, Dec. 2002.
[11]I. Dinur and K. Nissim, "Revealing Information while Preser-ving Privacy," Proc. ACM Symp. Principles of Database Systems (PODS '03), 2003.
[12]C. Dwork, "A Firm Foundation for Private Data Analysis," Comm. ACM, vol. 54, no. 1, pp. 86-95, 2011.
[13]C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our Data Ourselves: Privacy via Distributed Noise Generation," Proc. 25th Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT '06), 2006.
[14]C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," Proc. Theory of Cryptography Conf. (TCC '06), 2006.
[15]A. Frank and A. Asuncion, UCI Machine LearningRepository, http://mlearn.ics.uci.edu/MLRepository.html, 2010.
[16]A. Friedman and A. Schuster, "Data Mining with Differential Privacy," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '10), 2010.
[17]B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, vol. 42, no. 4, pp. 1-53, June 2010.
[18]B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.
[21]O. Goldreich, Foundations of Cryptography, vol. 2, Cambridge Univ. Press, 2001.
[22]O. Goldreich, S. Micali, and A. Wigderson, "How to Play Any Mental Game—A Completeness Theorem for Protocols with Honest Majority," Proc. ACM Symp. Theory of Computing (STOC '87), 1987.
[23]M. Hay, V. Rastogi, G. Miklau, and D. Suciu, "Boosting the Accuracy of Differentially Private Histograms through Consis-tency," Proc. Int'l Conf. Very Large Data Bases (VLDB '10), 2010.
[24]A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino, "Private Record Matching Using Differential Privacy," Proc. Int'l Conf. Extending Database Technology (EDBT '10), 2010.
[25]V.S. Iyengar, "Transforming Data to Satisfy Privacy Constraints," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), 2002.

AUTHOR'S PROFILE

Author's Name : Gauri V.Sonawane
is a P.G. student of Computer Engineering at SITRC College of Engineering , Nasik under Savitribai Phule Pune University . She has completed her undergraduate Course of engineering from Savitribai Phule Pune University. Her areas of interest include Data Mining..

Author's Name : Prof. Naresh Thoutam
Completed his M.Tech M.Tech(CSE) from Walchand College Of Engineering ,Sangli.And B.Tech(CSE) from JNTU KAKINADA.Presently he is working as Assistant Professor at SITRC College of Engineering, Nasik, Maharashtra, India.