Available online at www.worldscientificnews.com

World Scientific News

WSN 45(2) (2016) 126-136 EISSN 2392-2192

Survey of Secure Two Parties Confidential Information Release in Vertical Partitioned Data

Dr. M. Newlin Rajkumar[1,a], **A. Vijayalakshmi**[2,b], **M. U. Shiny**[3,c]

[1]Assistant Professor, Department of CSE, Anna University Regional Campus Coimbatore, Tamil Nadu, India

[2,3]PG Scholar, Department of CSE, Anna University Regional Campus Coimbatore, Tamil Nadu, India

[a-c]E-mail address: newlin_rajkumar@yahoo.co.in , vijimecse4@gmail.com , shinysusma05@gmail.com

ABSTRACT

To securely give person-exact fragile data from two data providers, whereby the mutual data maintains the necessary information for behind data mining tasks. Secured information distributed locations the problem of discovery delicate information when mining for useful data. In this paper, we address the problem of private data publishing, where personal uniqueness for the similar arrangement of people is detained by two parties. Differential privacy is a detailed security show that makes no doubt around an adversary's experience background knowledge. A differentially-private module ensures that the probability of any output (discharged information) is immediately as likely from all about the same information sets and therefore ensures that all outputs are merciless to any singular's data. As it were, a singular's privacy is not at risk in light of the interest in the data set. Specifically, we show a control for differentially private data discharge for vertically-distributed data between two parties in the semi-genuine adversary model. We first present a two-party convention for the exponential mechanism .This convention can be used as a sub convention by some other computation that requires the exponential component in a distributed setting. Likewise, we propose a two-party algorithm that discharges Differentially-private information in a secure way as per the significance of secure multiparty computation.

***Keywords*:** Differential privacy; Two party algorithm; Exponential Mechanism

1. INRODUCTION

Database is a good communication and storing system. Each database is owned by a exacting independent entity, for example, medical data, income data, financial data, and census data are using in several field. These disseminated data can be included to better data analysis for making better decisions and providing high-quality services. For example, integrated data can be improved spatial research, customer service, etc. The data integration between independent entities should be conducted in such a technique that no more details than necessary is exposed between the participating entities.

New knowledge that outcome from the integration process should not be tainted by adversaries to show susceptible information that was not available before the data integration. The proposed algorithm to securely integrate person-specific susceptible data from three data providers, the integrated data still hold the necessary information for supporting data mining tasks. Some kind of applications falls in to vertically partitioned applications where with same id have different set of attributes stored at different sites. It also Provide the susceptible information in Encrypted Format.

For example: All the client details stored in the Database of the Service Provider. Once the client creates an account, they are to login into their account from the Service Provider. Based on the client's request, the Service Provider will process the client requested Job and respond to them. The client information will be stored in the Database of the Company Service Provider. Company server will control the large amount of data in their Data Storage. The Company Service provider will maintain the all the client information to verify when they want to login into their account. The Company Server will transmit the client requested job to the any of the Queue to process the client requested Job. The Request of all the clients will process by Company Server will establish connection between them. We are going to create an client Interface Frame. Also the Company Service Provider will send the client Job request to the Queues in Fist in First out manner.

The Bank Service provider will maintain the all the client information to verify when they want to login into their account. The client information will be stored in the Database of the Bank Service Provider. Bank Service Provider will hold information about the user in their Data Storage. To converse with the Client and with the other modules of the Company server, the Bank Server will create connection between them. A idea of merged data that is in a company or organization will sustain the employee information both private and public data is implemented.

The employee may contain private data like employee id, name , salary and the loan applied and public data like email id, address and phone number. But more private information like bank account number and pin number are not reveled form the company. Therefore we merge the private and public information into one new table. The two party verification are done by both bank and company to list the log of the employee whether he is eligible to take up loan. so the verification by bank through the company and company will provide a set of information it will be validated by the both bank and company by using scheme of two party authentication.

2. LITERATURE SURVEY

2. 1. Data confidentiality through optimal K-annonymization

R. Agrawal and R. Srikant says about data sharing across private databases proposed the method for determining an K-optimal anonymization of a given dataset. An ideal anonymization is one which perturbs the input dataset as little as is important to accomplish k-anonymity, where "as little as is necessary" is typically quantified by a given cost metric. Ability to calculate optimal anonymizations is check the impacts of various coding techniques and problem variations on anonymizations quality. It allows good quality the effectiveness of stochastic or other non-optimal methods.

R. J. Bayardo and Agrawal discuss Data privacy through optimal k-anonymization has been planned to reduce the threat of this type of attack. The main objective of k anonymization is to protect the privacy of the singular's to whom the data pertains. But, subject to this limitation, it is important that the released data remain as "useful" as possible. Several recoding models have been proposed in the survey for k-anonymization, and often the "quality" of the available data is dictated by the model that is used.

2. 2. l - diversity

Machanavajjhala A, Kifer Gehrke D and Venkitasubramaniam M considered anonymity by the l-diversity corresponds to some notion of ambiguity of linking a QID to a particular sensitive value. Wang planned to bound the assurance of inferring a particular susceptible value using one or more confidentiality templates specified by the data supplier. Wong proposed some overview methods to concurrently achieve k-anonymity and bound the confidence.

2. 3. t - Closeness

Privacy away from k-Anonymity and l- Diversity as k-anonymity protects against individuality discovery, it does not supply sufficient protection against quality discovery. The concept of l-diversity attempts to resolve this problem by requiring that each equivalence class has at least l well-represented values for each susceptible quality. Show that l-diversity has a number of boundaries and have proposed a novel privacy idea called t-closeness, which requires that the distribution of a susceptible quality 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). One key uniqueness of our approach is that we isolate the data gain an observer can get from a released data table into two parts: that about all population in the discharged data and that about specific singulars. This enables us to boundary only the second kind of information gain.

Multiple susceptible attributes present additional challenges. Assume we have two susceptible attributes U and V. One can consider the two attributes separately, i.e., an equivalence class E has t-closeness in the event that E has t-closeness concerning both U and V. Another approach is to regard as the joint distribution of the two attributes. To utilize this approach, one has to choose the ground distance between pairs of susceptible attribute values. A basic formula for computing EMD may be difficult to derive, and the relationship between t and the level of privacy becomes more complicated.

2. 4. Anonymizing Classification Data for Privacy Preservation

Anonymizing Classification Data for Privacy Preservation consider the problem of ensuring an singular's anonymity while releasing person specific data for classification analysis. Specify that the before ideal k-anonymization based on a closed-form cost metric does not address the classification requirement. Our approach is depends on two observations specific to classification: Information specific to singulars tends to be over appropriate, thus of little utility, to classification; even if a masking operation eliminates some useful Classification structures, another structures in the data emerge to help. So, not all data items are uniformly useful for classification and less useful data things give the space for anonymizing the data without compromising the utility.

2. 5. The Limits of Two-Party Differential Privacy

It study about differential privacy in a disseminated setting where two parties would like to perform analysis of their combined data while preserving privacy for both datasets. It outcome involve almost tight lower bounds on the correctness of such data analyses, both for specific natural functions (such as Hamming distance) and in general. The bounds expose a sharp contrast between the two-party site and the simpler client-server site (where privacy guarantees are one-sided). In addition, those bounds reveal a dramatic gap between the correctness that can be obtained by differentially private data analysis versus the correctness obtainable when privacy is relaxed to a computational deviation of differential privacy.

2. 6. Two-party algorithm

In this paper, Distributed Differentially-private anonymization algorithm based on Generalization (DistDiffGen) for two parties. The algorithm first generalizes the raw data and then adds noise to achieve differential privacy. The general thought is to anonymize the raw data by a sequence of specializations starting from the topmost common state. A specialization, written v → child (v), where child (v) denotes the set of child values of v, replaces the parent value v with child values. The specialization method can be viewed as pushing the "cut" of each taxonomy tree downwards. A cut of the taxonomy tree for an attribute A^{pr}_i, denoted by Cut_i, contains exactly one value on each root-to-leaf path. The specialization starts from the uppermost cut and pushes down the cut iteratively by specializing a value in the current cut. Algorithm is executed by the party P1 (same for the party P2) and can be summarized as follows:

- Generalizing raw data
- Computing the Count
- Computing the Noisy Count

2. 6. 1. Generalizing raw data

Candidates are preferred based on their score values, and different utility functions can be used to determine the scores of the candidates. Once a victor candidate is determined, both parties specialize the winner on generalize data table by part their records into child partitions agreeing to the provided taxonomy trees.

2. 6. 2. Computing the Count

For each leaf node in the resulted generalize data table from the previous step, parties require computing the true count before adding noise. It uses the Secure Scalar Product Protocol for counting.

2. 6. 3. Computing the Noisy Count

To compute the overall noisy count, the main party produces two Gaussian random variables, that distributed values for the mean and the variance.

2. 7. Two-party differentially confedential information release algorithm

Algorithm. Two-Party Algorithm (DistDiffGen).

Input: Raw data set D1, privacy budget ∈,

and number of specializations h

Output: Anonymized data set ^D

```
Initialize Dgwith one record containing top most values;
Initialize Cuti to include the topmost value;
$\in' = \in 2( Anpr + 2h)$
Determine the split value for each vn∈∪Cuti withprobability∝exp $\in' 2\Delta uu\ D,n$ ;
Compute the score $\forall v \in$∪Cut$i$
for l= 1 to h do
    Determine the winner candidate $w$by Algorithm (DistExp);
    if w is local then
        Specialize $w$ on Dg;
        Replace $w$ with child ($w$) in the local copy of ∪Cuti;
        Instruct P2 to specialize and update ∪Cuti;
        Determine the split value for each new vn∈∪Cuti with probability ∝exp
            $\in' 2\Delta uu\ D,n$ ;
        Compute the score for each new $v\in$∪Cut$i$
```

```
else
Wait for the instruction from P2;
Specialize w and update ∪Cuti using the instruction;
end if
end for
for each leaf node of Dg does
Execute the SSPP Protocol to compute the shares C1And C2 of the true countC;
Generate two Gaussian random variables $Yi \sim N\ (0,\sqrt{1}\in)$ for $i \in 1,2$ ;
Compute $X1=C1+Y12-Y22$
Exchange X1 with P2 to compute (C+Lap (2/∈));
end for
return each leaf node with count (C+Lap(2/∈)) ;
```

3. RELATED TECHNIQUES

3. 1. Interactive Versus Non-Interactive

In an intellectual system, can posture questions through a private instrument, and a database owner answers these questions consequently. In a no interactive system, a database owner first anonymizes the basic information and after that releases the anonymzed version for information examination. Once the data is distributed, the information owner has no further control over the disseminated data. This methodology is otherwise called privacy preserving data publishing (PPDP).

3. 2. Single Versus Multiparty

Information or data may be claimed by a single party or by multiple parties. In the appropriated (multiparty) situation, information owners need to achieve the same tasks as single gatherings on their integrated information without imparting their information to others. Our proposed algorithm addresses the distributed and non-interactive scenario.

3. 3. Single Party Scenario

It provide an overview of some related anonymization algorithms. Many algorithms is proposed to preserve privacy, but only a few have measured the goal for classification analysis. It presented the secrecy problem for classification and proposed a genetic algorithmic solution. Using the top down specialization (TDS) approach to simplify a data table. The differential privacy concentrates on the interactive setting with the goal of dropping the magnitude of the added noise, releasing data mining results, or determining the feasibility and infeasibility solution of differentially-private mechanisms. It address the problem of non-interactive data release only the single-party scenario. Therefore, these techniques do not satisfy the confidentiality requirement of our data integration function for the economic industry.

3. 4. Distributed Interactive Approach

Distributed Interactive Approach is referred to as privacy preserving disseminated data mining (PPDDM). In PPDDM, some data owners want to compute a function based on their inputs without sharing their data with others. For example, several hospitals using data mining model for disease based on patients' medical history without distribution their data with each other. In recent years, dissimilar protocols have been proposed for different data mining tasks including association rule mining, clustering and classification. Compared to an interactive approach and a non-interactive approach provide greater suppleness because data recipients can perform their required analysis and data exploration, such as data mining patterns in a specific group of records, visualizing the transactions containing a specific model.

3. 5. Distributed Non-Interactive Approach

This approach allows anonymizing data from different sources for release the data without sensational the perceptive information. This algorithm to securely integrate horizontally splitted data from multiple data owners without disclosing data from one party to another. Distributed algorithm to add horizontally-partitioned high dimensional health care data. Anonymization algorithm to integrate data from multiple data owners. Interactive and non-Interactive are the only two methods that generate an integrated anonymous table for vertically partitioned data. The both methods adopt k-anonymity, or its extensions, as the underlying privacy principle and, therefore, both are susceptible to the recently discovered privacy attacks.

4. MECHANISM OF PRIVATE DATA RELEASE FOR VERTICALLY PARTITIONED DATA

The algorithm for differentially-private data release for vertically partitioned data between two parties. Mohammed has been proposed the single-party algorithm for differential confidentiality as a basis and extends it to the two-party setting.

We using another algorithm satisfy the security description of the semi-honest adversary model. In this method, parties follow the algorithm but may try to deduce extra information from the received messages. Whenever amid the execution of the algorithm, no party should learn more information about the other part's information than what is found in the last integrated table, which is differentially-private.

The main part of our paper can be summarized as follows: We introduce a two-party protocol for the exponential mechanism. We use this protocol as a sub protocol of our main algorithm, and it can likewise be used by any other algorithm that uses the exponential mechanism in a distributed setting. Differential privacy provides strong statistical privacy guarantees for certain types of queries, even in worst-case scenarios.

In this method having three mechanisms as follows:

- Vertically partitioning on given dataset
- Exponential mechanism
- Privacy for the release scattered data

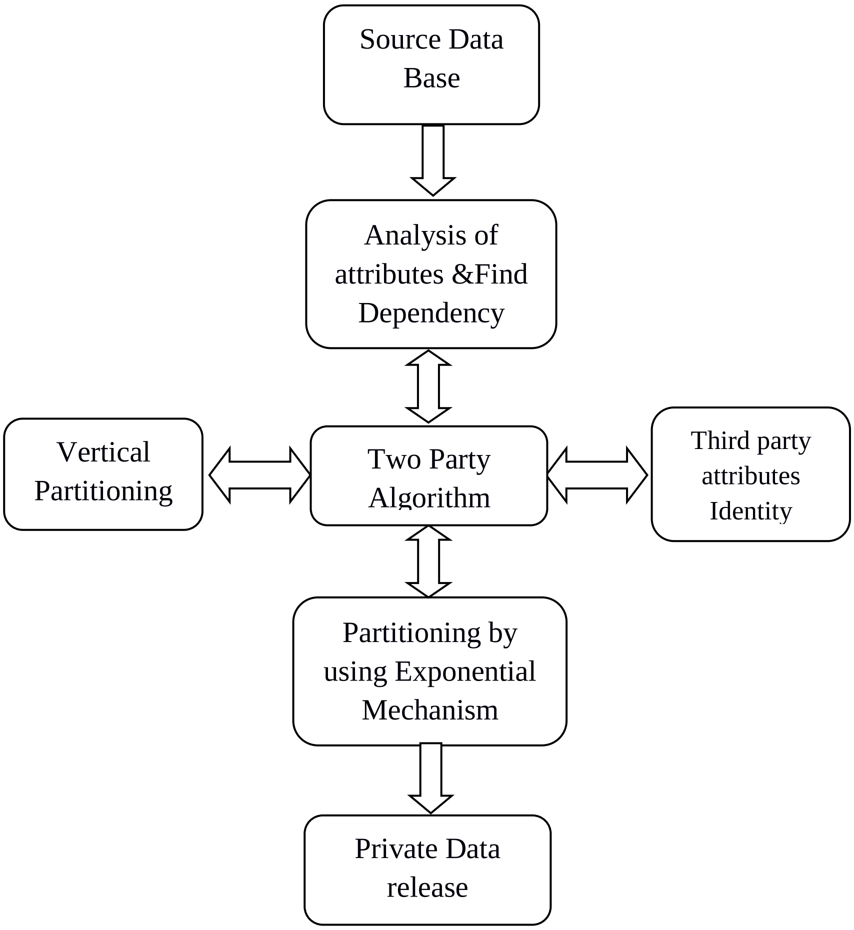
Source Data Base
Analysis of attributes &Find Dependency
Vertical Partitioning
Two Party Algorithm
Third party attributes Identity
Partitioning by using Exponential Mechanism
Private Data release

Figure 1. System Architecture

Vertically partitioning on given dataset

We partitioning our source data that is the bank data then we have two entities the first entity will hold the profile data which not that much sensitive and the other entity hold the sensitive data.

Exponential mechanism

In this method of the exponential partitioning mechanism algorithm we integrated the data which is generated from the vertically partitioned data that satisfies the requirements of the differential privacy.

Privacy for the release scattered data

In last method of the privacy for the release scattered data we are providing security for the data which is release from the distributed framework.

5. CONCLUSION

We using that the vertically partitioned data by using of the exponential mechanism ensures that the algorithm convince by the differential privacy model and also secures the data in the scattered framework. Our second algorithm guarantees that the private data which is released from the scattered framework are secured. It provides improved data utility than the single-party algorithm and the distributed k-anonymity algorithm.

6. OVERALL IDEA OF THE PAPER

Information between two parties where incorporated by using common identifier such as name, employee id. Incorporated data is pre-processed ie.. Removing all the unambiguous identifiers such as name, age, etc.. but there may be a survival of fake identifiers which may lead to link attack. Incorporated data gets generalized to hide the susceptible details. Owner of the data generalizes the details by assuming some of the field as susceptible. Thus security is satisfied statistically. A method is to provide dynamic security called differential privacy which does not assume about adversaries background knowledge.

References

[1] Barak B., Chaudhuri K., Dwork C., Kale, McSherry F. and Talwar K. (2007), ‘Privacy Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release’, Proc. ACM Symp. Principles of Database Systems (PODS ’07).

[2] Bayardo R., and Agrawal R (2005), ‘Data privacy throughoptimal k-anonymization’. In Proceedings of the IEEE International Conference on Data Engineering (ICDE).

[3] Chaudhuri K., Monteleoni C. and Sarwate A (2011), ‘Differentially private empirical risk minimization’. *Journal of Machine Learning Research* (JMLR), 12; 1069-1109.

[4] Chaudhuri K., Sarwate A D. and Sinha K (2012), ‘Near-optimal differentially private principal components’, In Proceedings of the Conference on Neural Information Processing Systems.

[5] Clifton C.,Kantarcioglu M., Vaidya J., Lin X., and Zhu M Y (2002), 'Tools for privacy preserving distributed data mining', ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD) *Explorations Newsletter,* 4(2); 28-34.

[6] Dwork C., McSherry F., Nissim K., and Smith A. (2006), 'Calibrating Noise to Sensitivity in Private Data Analysis', Proc. Theory of Cryptography Conf. (TCC 06).

[7] Dwork C., Kenthapadi K., McSherry F., Mironov I., and Naor M. (2006), 'Our Data Ourselves: Privacy via Distributed Noise Generation', Proc. 25th Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT '06).

[8] Dwork C. (2011), 'A Firm Foundation for Private Data Analysis,' *Comm. ACM,* Vol. 54, No. 1, pp. 86-95.

[9] Fung B.C.M.., Wang, K.., Chen, R.., Yu, P.S (2007), 'Privacy-preserving data publishing', A survey of recent developments. *ACM Computing Surveys,* Vol. 42, No. 4, pp. 1-53.

[10] Fung B.C.M., Wang, K., Yu, P.S (2007), 'Anonymizing classification data for privacy preservation', *IEEE Transaction on Knowledge and Data Engineering* (TKDE), Vol. 19, No. 5, pp. 711-725.

[11] Fung B C M., Wang K., Chen R. and Yu P. S. (2010), 'Privacy-Preserving Data Publishing: A Survey of Recent Developments', *ACM Computing Surveys,* Volume. 42, No. 4, pp. 1-53.

[12] Jiang W and Clifton C (2006), 'A Secure Distributed Framework for Achieving k-Anonymity,' *Very Large Data Bases J.,* Vol. 15, No. 4, pp. 316-333.

[13] LeFevre, K.., DeWitt, D.J., Ramakrishnan, R (2006). 'Mondrian multidimensional k-anonymity'. In Proceedings of the IEEE International Conference on Data Engineering (ICDE).

[14] Lindell Y and Pinkas B(2002), 'Privacy Preserving Data Mining', *J. Cryptology,* vol. 15, No. 3, pp. 177-206.

[15] Machanavajjhala A., Kifer D., Gehrke J., Venkitasubramaniam, M. (2007), 'ℓ-diversity: Privacy beyond k-anonymity', ACM Transactions on Knowledge Discovery from Data (TKDD).

[16] Mohammed N., Alhadidi D., Fung B C M (2014), 'Secure Two-Party Differentially Private Data Release for Vertically Partitioned Data', Proc. *IEEE Transaction On Dependable And Secure Computing* Vol. 11, No.1, pp. 59-70.

[17] Mohammed N., Chen R., Fung B C M and Yu P S (2011), 'Differentially Private Data Release for Data Mining', Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '11).

[18] Mohammed N., Fung B C M., and Debbabi M (2011), 'Anonymity Meets Game Theory: Secure Data Integration with Malicious Participants', *Very Large Data Bases J.,* Vol. 20, No. 4, pp. 567-588.

[19] Narayan A. and Haeberlen A. (2012), 'DJoin: Differentially Private Join Queries over Distributed Databases', Proc. 10th USENIX Conf. Operating Systems Design and Implementation (OSDI '12).

[20] Vaidya J and Clifton C (2002), 'Privacy Preserving Association Rule Mining in Vertically Partitioned Data', Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02).

[21] Vaidya J and Clifton C. (2003), 'Privacy-Preserving k-Means Clustering over Vertically Partitioned Data,' Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '03).

(Received 12 March 2016; accepted 26 March 2016)