ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 15, March 2016

Privacy Preserving Data Publishing Using Sliding Window Query For Data Stream

B.Vinodhini[1]
M.E., Computer Science Student,
M.I.E.T Engineering College, Trichy,
vinobaskar04@gmail.com

S.Mohana[2]
Assistant Professor, Department of CSE,
M.I.E.T Engineering College, Trichy
mohana.p3@gmail.com

G. Grace Ranjithm[3]
M.E., Computer Science Student,
M.I.E.T Engineering College, Trichy
grace.georgekumar@gmail.com

Abstract - Access Control Mechanisms (ACM) and Privacy Protection Mechanisms (PPM) protect sensitive information from unauthorized users. The access control for a data stream allows us to access the database under Role Based Access Control (RBAC) policy that satisfies an authorized predicate sliding-window query. A PPM can use generalization of relational data to anonymize and satisfy privacy requirements such as k-anonymity and l-diversity. Generalization can reduce the delay in publishing of stream data through imprecision bound.Delay in sharing the data stream leads to false negatives i.e., it satisfies the sliding-window query at the time instance of query evaluation. Access control policy is defined for the imprecision bound for each query. To optimize the delay in publishing the stream data is the most difficult task for Privacy Protection Mechanism. To overcome this challenge the imprecision bound for maximum number of queries has to be satisfied. The Precision-bounded Access Control for Privacy-Preserving data strEams (PACE) problem generate an anonymized data stream such that the average number of times the query imprecision bound is violated over a given time period.

Keywords – k-Anonymity, RBAC, data stream, privacy.

I. INTRODUCTION

DATA Stream Management System (DSMS) is a collection of database that allows us to access transactional datasuch as internet traffic, health monitoring, and sensor networks. A data Stream is an unbound sequence of data elements. Collections of data refer as Microdata that contain sensitive information of an individual user. If the sensitive information of the data stream is not privacy protected, then the privacy of a person can be compromised even in the presence of access control. Hence, access control mechanisms for datastreams ensure that only the authorized parts of the streamsare available to each user. Objects are protected and accessed by using queries or views of data stream. ACM that allows user to access the data based on Role Based Access Control Policy. RBAC that defines permission for each object based on roles. Privacy Protection Mechanism is used to preserve the privacy of data respondent. Privacy can be attained by either Suppression or Generalization or both. Privacy technique is mainly used to anonymize and satisfy the privacy requirement i.e., k-anonymity and l-diversity subjected to identity disclosure of authorized users.

In this paper, it mainly focus on preserving the privacy of individual information from unauthorized user. Here, individual information refers as microdata that containName, Age, Gender, Zipcode, Address, and Salary. Anonymization is the process that removes identifiers such as Name, Address, Social Security number which is publicly available data such as Voter Registration data, After removing the identifiers, it classify the table as Sensitive attributes (Medical condition, Salary) and Non-Sensitive attributes (Age, Zipcode, Nationality). Anonymization is performed to hide the sensitive information of a respondent,The attribute values in data stream tuples can be generalizedto maintain given privacy requirements.Generalizationof relational data attributes introduces imprecision boundfor the query results. This imprecision bound can be used to reduce delay in publishing of stream data.

A. Related Works

Let us consider the previous work of this paper is related to privacy preserving publishing of data streams in which precision-bounded access control and privacy has not been investigated before. Nehme et al. propose security punctuation-based access control framework for data streams. Security punctuation is a predicate that defines access to stream data and is created by the user generating stream data. The subjects are assigned roles on the server and execute authorized queries on the incoming data stream. Role-based access control for data streams has been proposed. In their framework, thereare two types of temporal constraints.First is the interval constraint during which the role can access stream data.Second is the window constraint that limits access to the data stream for each role according to the authorized view defined by the sliding-window query predicate.Cao et al. have proposed CASTLE for continuously anonymizing data streams in which it extend the definition ofk-anonymity for data streams and propose a clustering algorithm that publishes anonymized clusters before maximum delay occur. To overcome this, Zhou et al. Proposed a randomized-algorithm based on the R-tree.

Dwork et al. have proposed differential privacy for data streams considering a single aggregate query. Cao et al.extend the model to sliding-window queries over binary data streams. Differential privacy is achieved by adding random noise to original query results and offers better privacy guarantees than generalization, similarly syntactic anonymization techniques (e.g., generalization) providebetter precision.

B. Contributions

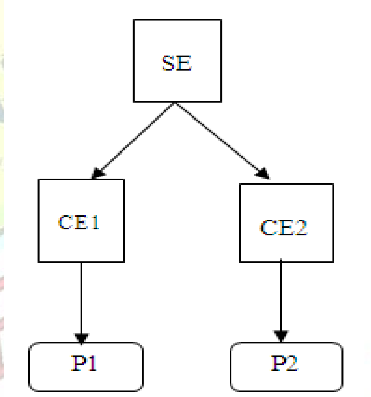
The contribution of the paper is to protect the sensitive information of the individual respondent from the unauthorized users. Hence, formulate the Precision-Bounded Access Control for Data strEams (PACE) problem and gives the probabilistic analysis for query bound violation. In the Query Bound Violation method is to access the user data by using query predicate and detect number of times the query imprecision bound is violated over a given time period.

Privacy Preserving Data Stream is achieved by anonymization method i.e., with the help of Generalization and Suppression technique. Both techniques are used to anonymize the user information.In earlier method, they use k-anonymity and l-diversity for anonymization but this technique does not protect the information in a secure way. In this technique an unauthorized user can easily access the sensitive information through publicly available information of that user. After anonymization, partition the data with respect to age and Zipcode so that receiver can easily retrieve the data from the database. The algorithm Total Imprecision Minimization (TIM) heuristic for multi-dimensional partitioning of stream data to satisfies imprecision bound for predicate sliding-window queries. TIM avoids overlapping of user data in query predicates by using R^+ tree based anonymization algorithm in which leaf nodes and intermediate nodes are non-overlapping. Finally, the user information has to be protected in an secure way.

II. ROLE BASED ACCESS CONTROL

An RBAC allows defining permission on object based on roles. RBAC policy ρis a tuple <U,R, P, UA,PA, RH>, where U is a set of users, R is a set ofRoles, P is a set of Permissions, RH is a Role Hierarchy denotesa partial order on (R×R) roles, UA is a user-to-role (U×R) assignment relation,and PA is a role-to-permission (R×P) assignment relation. when a userassigned a role to executes a query, the tuples satisfying the query predicate. Permissions under P are the sliding-window query predicates that define the authorized view of the data stream.

Role	Designation
SE	State Epidemiologist
CE1	Country 1 Epidemiologist
CE2	Country 2 Epidemiologist



Permission	Authorized Query Predicate View
P1	Location=CE1^Age=5-50^Syndrome=Influenza
P2	Location=CE2^Age=5-50^Syndrome=Influenza

Fig.1 Role Based Access Control Policy

Motivating Scenario: Syndromic surveillance systems are used to detect and monitor the threads to public health at the state and country level. The department of health in a state collect the emergency symptoms from country hospital in an hourly basis and update the data, classify the data into syndromecategoriesaccording to department of health. An access control policy allows the roles to access the tuples under the authorized predicate i.e., in Fig1 Role CE1 can view tuples under permission P1 over the data stream in a 24hour basis and sliding the window for every 4hours (slide=4hours) in which query can be executed. Permission under an access control policy assures only the authorized view of the data stream is available to each user. Anonymization adds False positives (FP) to the authorized view and the privacy can be achieved by delay in data stream similarly delay adds False Negatives (FN) for the authorized views. The total imprecision bound for query predicate is the sum of false positives and false negatives so that user can predict the accurate information of an individual user.

***False Positive Tuples*:** A tuple is a false-positive when it does not satisfy the sliding-window query predicate at the time instance of query evaluation but is included in the query result as the equivalence class in $T^p[i]$ that contains the tuple overlaps the query predicate.

$$FP_{Qj}[i] = |Q_j(T^p[i])| - |\, Q_j(\, T[i] - T^h[i]\,)\, |, \text{ where}$$

$$|\, Q_j(T^p[i]\,)\, | = \Sigma_{EC(overlaps)Qj} |\, EC\, |$$

False Negative Tuples: A tuple is a false-negative when it satisfies the sliding-window query at the time instance of query evaluation but is not included in the query result due to being put on hold.

$$FN_{Qj}[i] = |\, Q_j(T^h[i]\,)\, |$$

If the publishing delay is increased, the number of false positives reduces because equivalence classes with less imprecision can be formed while at the same time the numberof false-negatives increases.

Sliding-Window Query Imprecision Bound: Query imprecision is defined as the total sum of false-positives and false-negatives for a sliding-window query method to evaluate an anonymized stream $T^p[i]$ at any given time instance i.The imprecision for query $Q_j[i]$ evaluated at time instance i is denoted by $imp_{Qj}[i]$ and is equal to the sum of false-positives and false-negatives.

$$imp_{Qj}[i] = FP_{Qj}[i] + FN_{Qj}[i]$$

Here, $imp_{Qj}[i]$ is a utility measure for a given sliding-window query captures two types of information loss; loss due to generalization (in terms of false positives) and loss due to publishing delay (in terms of false negatives).

III. SYSTEM MODEL

A precision bound access control framework for privacy-preserving data streams has been proposed by Access Control Mechanism and data are protected under Privacy Protection mechanism. In the PPM, it mainly focus on protecting the user information. Here, admin collect the micro data of the patient from various hospital and placed it into the database. After collecting, partitioning the data set in to two sets one is anonymous data another is sensitive data. Anonymization is done by using privacy requirements. In this way, Privacy Protection Mechanism is to protect the sensitive information. Access Control Mechanism gives permission to the user to access the authorized information. Once permission is permitted then imprecision bound has to set the boundary for the query predicate and check whether the user has the rights to access to data with the help of reference monitor. If user is an authorized user then the user can view and access the sensitive information so that Precision gets improved and time complexity is reduced in the system.

A. Microdata Creation Module

Collect patient data through any hospital or internet and import the data to our application. Collection of data referred to as Micro-Data. Micro-Data contains details of patient such as ID, Name, Age, Address, Zipcode, Nationality and Salary. These data can be placed as Identifiers, Quasi identifier and Sensitive Information. Each patient has a unique identifier, Quasi Identifier includes data of Name, Age, Address, Zipcode, Nationality whereas sensitive attributes includes patient disease or salary. After collecting the microdata, data can be imported, viewed and accessed by administrator. Here, Micro-Data may be a voter list or patient record.

B. Anonymization Creation Module

After collecting the Micro-Data, data get anonymized. In the anonymization process, Identifiers are removed first then divide the attributes into two groups: one is sensitive attributes (consisting only of medical condition) and the non-sensitive attributes are Zipcode, Age and Nationality.

An attributes is marked sensitive and not allowed to discover the value of attributes for any individual dataset and others are referred as non-sensitive. In 4-anonymous table derived from generalization and suppression. Here "*" denotes a suppressed value i.e., "Zipcode = 1485*" means that the Zipcode is in range [14850-14859] and generalization denotes range i.e., "age = [15-30].

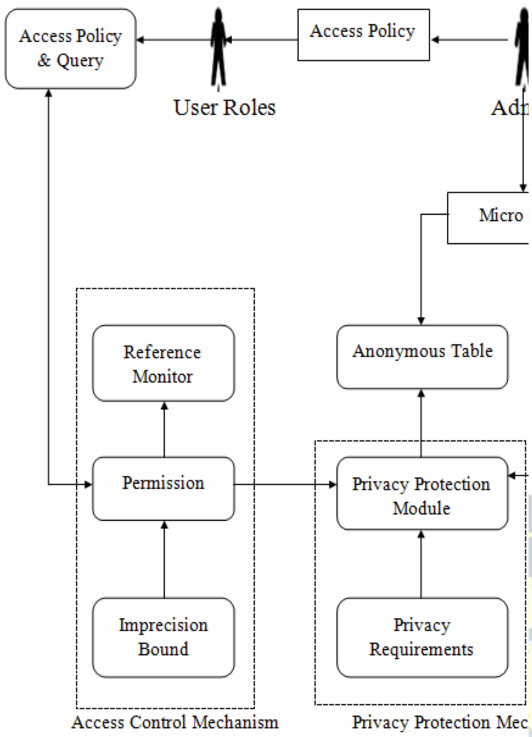


Fig.2 System Architecture

C. Data Partitioning Module

After collecting the Micro-Data, the data can be converted into anonymous table where the identifier has to be removed, quasi identifier and sensitive information has to be processed. After that, initialize the partition according to age or Zipcode, then split the data using sliding-window query method. Partitioned table can be verified if any overlapping occur in the data set i.e., if one of the patient occur in the query partitioning P1 whereas the same user also occur in another partitioning P2 can be viewed by Partitioning Graph. If overlapping occur that can be overcome by the Expected False-Positives for a leaf node Partitioning (EFPP) is defined as the sum of false-positives for all queries resulting from Partition P is published at the current time instances. A sliding–window query adds false-negatives that are evaluated in the next time instances.

D. Data Stream Module

In the data stream module, access control policy is defined; that allows the roles to access the tuples under the authorized predicate, e.g., Role CE1 can access tuples under Permission P1.Once the query predicate is defined the respondent has verified whether the predicate is available in imprecision bound and user has rights to access that data. After successfully convert to anonymous table and also assign a role based access mechanism over internet application to publish the anonymous table to other research centre over the world. These data are fully secured and then published to internet.

E. Data Access Module

The research centre to access the anonymous data through our access keys or our SSN number over government server because data published to the internet is in encrypted form. Only the authorized user know the access key to access the data whereas unauthorized user doesn't access the authorized information so that data are kept more securely and reduce time complexity.

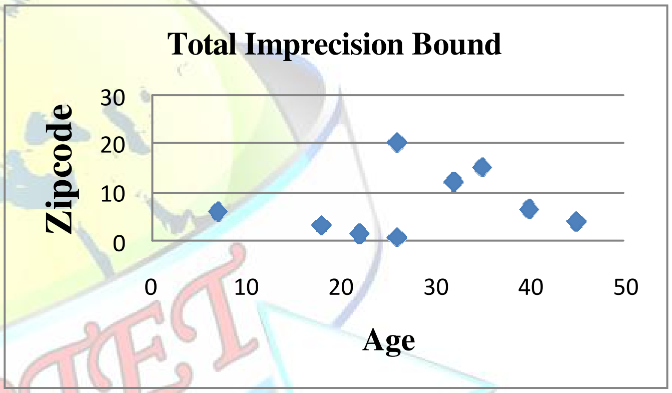


Fig.3 TIM for Query Predicate

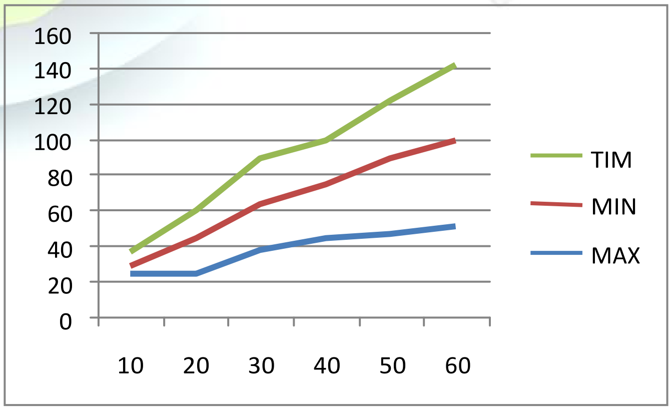


Fig.4 TIM with R^+ tree for dataset

IV. CONCLUSION

Precision-bounded access control for privacy preserving data streams has been proposed. The access control mechanism defines the permitted view of the data stream along with the required precision. The privacy protection mechanism applies generalization to the stream data such that the privacy requirement is met and imprecision bound for the maximum number of sliding-window queries is satisfied. An algorithm has been proposed to minimize the total imprecision bound and experiment has been performed to reduce the time complexity.

V. FUTURE ENHANCEMENT

In future work, extend the access control enforcement to the *Enclosed semantics* and also extend the privacy as *differential privacy* model for sliding window queries over the binary data stream to the relational data streams.

REFERENCES

[1] S.Chaudhuri, T. Dutta, and S. Sudarshan, “Fine Grained Authorization through Predicated Grants,” Proc. IEEE 23rd Int’l Conf. Data Eng.,pp.2007.

[2] D. Ferraiolo, R. Sandhu, S. Gavrila, D. Kuhn, and R, Chandramouli, “Proposed NIST standard for role-based access control,”ACMTrans.Inf.Syst. Security, vol. 4, no. 3, pp. 224–274,2001.

[3] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, “A framework for efficient data anonymization under privacy andaccuracy constraints, ACMTrans. DatabaseSyst., vol. 34, no. 2, p. 2009.

[4] K. LeFevre, R. Agrawa, V. Ercegovac, R. Ramakrishnan, Y. Xu, and D. DeWitt, “Limiting Disclosure in Hippocratic Databases,” Proc. 30th Int’l Conf.Very Large Data Bases, pp. 108-119.2004

[5] K. LeFevre, D. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” in Proc. 22nd Int. Conf. Data Eng., p. 25.2006

[6] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” ACM Trans. Knowl. Discov.Data, vol. 1, no. 1, p. 3.2007.

[7] R. Nehme, E. Rundensteiner, and E. Bertino, “A security punctuation framework for enforcing access control on streaming data,” in Proc.IEEE 24th Int. Conf.Data Eng., pp. 406–415,2008.

[8] Z. Pervaiz, W.G. Aref, A. Ghafoor, and N. Prabhu, “Accuracy constrained privacy-preserving access control mechanism for relational data,” IEEETrans. Knowl. Data Eng., vol. 26, no. 4, pp. Apr.2014

[9] N. Li. W. Qardaji, and D. Su, “Provably Private Data Anonymization: Or, k-Anonymity Meets Differential Privacy,” preprint arxiv:1101.2604.2011

[10] B. Zhou, Y. Han, J. Pei, B. Jiang, Y. Tao, and Y. Jia, “Continuous privacy preserving publishing of data streams,” in Proc. 12th Int. Conf. Extending Database Technol.: Adv. Database Technol., pp. 648–659.2009.

B.Vinodhini pursuing M.E Computer Science And Engineering degree in M.I.E.T Engineering College, Trichy in 2016 and received B.E Computer Science And Engineering degree from Anna University Of Technology, Trichy in 2014.
Email: vinobaskar04@gmail.com

S.Mohana, she has received B.E Computer Science And Engineering degree from SRM Easwari Engineering College, Chennai in 2001 and she received ME Computer Science And Engineering degree from PABCET, Trichy in 2008. Currently pursuing Ph.D in Anna University.
Email: mohana.p3@gmail.com

G. Grace Ranjitham pursuing M.E Computer Science And Engineering degree in M.I.E.T Engineering College, Trichy in 2016 and received B.E Computer Science And Engineering degree from Anjalai Ammal Mahalingam Engineering College, Kovilvenni in 2013.
Email: grace.georgekumar@gmail.com