Privacy Preserving and Access Control Mechanism for Health Database

Chandana R P, Shinu Mariam Koshy, Sreedevi S M, Sreeja S A (UG Students)
Lekshmy P L (Assistant Professor) Department of Computer Science and Engineering, LBS Institute of Technology for Women,
Thiruvananthapuram, Kerala

***Abstract*---- Access Control Mechanisms (ACM) is used to ensure that only authorized information is available to users. Privacy Protection Mechanism (PPM) uses suppression and generalization of relational data to anonymize and satisfy privacy needs. However, when sensitive information is shared and a Privacy Protection Mechanism (PPM) is not in place, an authorized user can still compromise the privacy of a person leading to identity disclosure. In this paper, we investigate privacy-preservation from the anonymity aspect. The sensitive information, even after the removal of identifying attributes, is still susceptible to linking attacks by the authorized users. The access control policies define selection predicates available to roles while the privacy requirement is to satisfy the k- anonymity or l-diversity. An additional constraint that needs to be satisfied by the PPM is the imprecision bound for each selection predicate. Top Down Selection Mondrian (TDSM) algorithm is used for query workload-based anonymization.**

.

Keywords*----*k-anonymity, l-diversity, imprecision bounds

I. INTRODUCTION

Access control mechanisms for databases allow queries only on the authorized part of the database. Predicate based fine-grained access control has further been proposed, where user authorization is limited to pre-defined predicates. Enforcement of access control and privacy policies has been studied. However, studying the interaction between the access control mechanisms and the privacy protection mechanisms has been missing. In this paper, we investigate privacy-preservation from the anonymity aspect. The sensitive information, even after the removal of identifying attributes, is still susceptible to linking attacks by the authorized users [2]. This problem has been studied extensively in the area of micro data publishing [3] and privacy definitions, e.g., k-anonymity [2], l-diversity [4] and variance diversity.

We use the concept of imprecision bound for each permission to define a threshold on the amount of imprecision that can be tolerated. Existing workload aware anonymization techniques [5] minimize the imprecision aggregate for all queries and the imprecision added to each permission/query in the anonymized micro data is not known. Making the privacy requirement more stringent results in additional imprecision for queries. The problem of satisfying accuracy constraints for individual permissions in a policy/workload has not been studied before. The heuristics proposed in this paper for accuracy-constrained privacy-preserving access control are also relevant in the context of workload aware anonymization. The anonymization for continuous data publishing has been studied in literature [3]. In this paper the focus is on a static relational table that is anonymized only once. To exemplify our approach, role-based access control is assumed. The concept of accuracy constraints for permissions can be applied to any privacy-preserving security policy, e.g., discretionary access control.

Example 1 (Motivating Scenario):

Threats to public health are studied and analyzed both at the state and federal level. Department of health in a state collects the emergency data from the county hospitals. Each daily update consists of a static instance that is classified into syndrome

categories by the department of health. Then, the surveillance data is anonymized and shared with departments of health at each county.

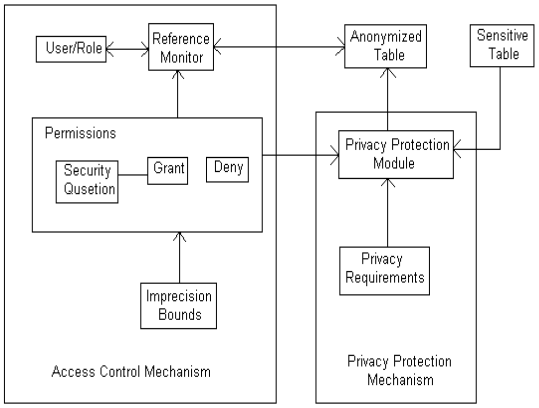

Fig. 1- System Architecture

II. BACKGROUND

This section focuses on role based access control and privacy definitions.

Given a relation T {A1, A2,...} where Ai is an attribute, T* is the anonymized version of the relation T. We assume that T is a static relational table. The attributes can be of the following types:

1) Identifier. These are attributes such as name, social security etc with which an individual can be uniquely identified. These attributes are completely removed from the anonymized relation.

2) Quasi-identifier (QI). These are attributes such as gender, zip code, birth date etc which can potentially identify an individual based on other information available to an adversary. QI attributes are generalized to satisfy the anonymity requirements.

3) Sensitive attribute. Attributes, e.g., disease or salary, that if associated to a unique individual will cause privacy breach.

III. RELATED WORK

Access control mechanisms for databases allow queries only on the authorized part of the database. Predicate based fine-grained access control has further been proposed, where user authorization is limited to pre-defined predicates. Enforcement of access control and privacy policies has been studied. However, studying the interaction between the access control mechanisms and the privacy protection mechanisms has been missing. Recently, Chaudhuri et al. have studied access control with privacy mechanisms. They use the definition of differential privacy whereby random noise is added to original query results to satisfy privacy constraints. However, they have not considered the accuracy constraints for permissions. We define the privacy requirement in terms of k-anonymity. It has been shown by Li et al. that after sampling, k-anonymity offers similar privacy guarantees as those of differential privacy. The proposed accuracy-constrained privacy preserving access control framework allows the access control administrator to specify imprecision constraints that the privacy protection mechanism is required to meet along with the privacy requirements.

The challenges of privacy-aware access control are similar to the problem of workload-aware anonymization. In our analysis of the related work, we focus on query-aware anonymization. For the state of the art in k-anonymity techniques and algorithms, we refer the reader to a recent survey paper [3]. Workload-aware anonymization is first studied by LeFevre et al. [5]. They have proposed the Selection Mondrian algorithm, which is a modification to the greedy multidimensional partitioning algorithm Mondrian. In their algorithm, based on the given query-workload, the greedy splitting heuristic minimizes the sum of imprecision for all queries. Iwuchukwu and Naughton have proposed an RB tree based anonymization algorithm [6]. The authors illustrate by experiments that anonymized data using biased RB tree based on the given query workload is more accurate for those queries than for an unbiased algorithm. Ghinita et al. have proposed algorithms based on space filling curves for k-anonymity and l-diversity. They also introduce the problem of accuracy-constrained anonymization for a given bound of acceptable information loss for each equivalence class. Similarly, Xiao et al. propose to add noise to queries according to the size of the queries in a given workload to satisfy differential privacy. However, bounds for query imprecision have not been considered. The existing literature on workload-

aware anonymization has a focus to minimize the overall imprecision for a given set of queries. However, anonymization with imprecision constraints for individual queries has not been studied before. We follow the imprecision definition of LeFevre et al. [5] and introduce the constraint of imprecision bound for each query in a given query workload.

IV. PROBLEM STATEMENT

Access Control Mechanisms (ACM) is used to ensure that only authorized information is available to users. Privacy Protection Mechanism (PPM) uses suppression and generalization of relational data to anonymize and satisfy privacy needs. Accuracy- constrained privacy-preserving access control framework is used to manage access control in relational database. The access control policies define selection predicates available to roles while the privacy requirement is to satisfy the k anonymity or l- diversity. Imprecision bound constraint is assigned for each selection predicate. K anonymous Partitioning with Imprecision Bounds (k- PIB) is used to estimate accuracy and privacy constraints. Role-based Access Control (RBAC) allows defining permissions on objects based on roles in an organization.

Top Down Selection Mondrian (TDSM) algorithm is used for query workload-based anonymization. The Top Down Selection Mondrian (TDSM) algorithm is constructed using greedy heuristics and k-d tree model. Query cuts are selected with minimum bounds in Top-Down Heuristic 1 algorithm (TDH1). The query bounds are updated as the partitions are added to the output in Top-Down Heuristic 2 algorithm (TDH2). The cost of reduced precision in the query results is used in Top-Down Heuristic 3 algorithm (TDH3). Repartitioning algorithm is used to reduce the total imprecision for the queries. The following problems are identified from the existing system.

- Static data based access control model
- Cell level access control is not supported
- Imprecision bound estimation is not optimized
- Fixed access control policy model

A. Data Partitioning

All the three algorithms based on greedy heuristics are proposed. All three algorithms are based on k-d tree construction. Starting with the whole tuple space the nodes in the k-d tree are recursively divided till the partition size is between k and 2k. The leaf nodes of the k-d tree are the output partitions that are mapped to equivalence classes [1]. Heuristic 1 and 2 have time complexity of O (d 2 *Q* n
2). Heuristic 3 is a modification over Heuristic 2 to have O (d|Q|nl gn) complexity, which is same as that of TDSM. The proposed query cut can also be used to split partitions using bottom- up (RB tree) techniques.

B. Anonymity Definitions

In this section, privacy definitions related to anonymity are introduced.

a) Equivalence Class (EC):

An equivalence class is a set of tuples having the same QI attribute values.

b) K-anonymity Property:

A table T satisfies the k-anonymity property if each equivalence class has k or more tuples [2]. K- Anonymity has been proposed as a mechanism for protecting privacy in micro data publishing, and numerous recoding "models" have been considered for achieving k anonymity. A number of organizations publish microdata for purposes such as demographic and public health research. In order to protect individual privacy, known identifiers (e.g., Name and Social Security Number) must be removed. In addition, this process must account for the possibility of combining certain other attributes with external data to uniquely identify individuals. For example, an individual might be "re-identified" by joining the released data with another (public) database on Age, Sex, and Zip code. K-anonymity has been proposed to reduce the risk of this type of attack. The primary goal of k-anonymization is to protect the privacy of the individuals to whom the data pertains. However, subject to this constraint, it is important that the released data remain as "useful" as possible.

c) Query Imprecision:

Query Imprecision is defined as the difference between the number of tuples returned by a query evaluated on an anonymized relation T*and the number of tuples for the same query on the original relation T.

d) Top Down Selection Mondrian:

The objective of Top Down Selection Mondrian (TDSM) algorithm is to minimize the total imprecision for all queries while the imprecision bounds for queries have not been considered. The anonymization for a given query workload with imprecision bounds has not investigated before to the best of our knowledge. We compare our results with TDSM in the experiments section. The algorithm presented is similar to the k-d tree construction. TDSM starts with the whole tuple space as one partition and then partitions are recursively divided till the time new partitions meet the privacy requirement. To divide a partition, two decisions need to be made

i) Choosing a split value along each dimension ii) Choosing a dimension along which to split.

In the TDSM algorithm [5], the split value is chosen along the median and then the dimension is selected along which the sum of imprecision for all queries is minimum. The partitions created by TDSM have dimensions along the median of the parent partition. A compaction procedure has been proposed in [6] where the created partitions are replaced by minimum bounding boxes. This step improves the precision of the anonymized table for any given query workload by reducing the overlapping partitions.

e) The k-PIB Problem:

The optimal k-anonymity problem has been shown to be NP-complete for suppression and generalization. The hardness result for k-PIB follows the construction that shows the hardness of k- anonymous multi-dimensional partitioning with the smallest average equivalence class size.

The cardinality of Query Qi is the sum of count values of tuples falling inside the query hyper- rectangle. The constant qv defines an upper bound for the number of queries that can violate the bounds.

V. HEURISTICS FOR PARTITIONING

This section includes three algorithms based on greedy heuristics are proposed. All three algorithms are based on k-d tree construction. Starting with the whole tuple space the nodes in the k-d tree are recursively divided till the partition size is between k and 2k. The leaf nodes of the k-d tree are the output partitions that are mapped to equivalence classes [1]. Heuristic 1 and 2 have time complexity of $O(Q|d^2|n^2)$

. Heuristic 3 is a modification over Heuristic 2 to have O (d|Q|n|gn) complexity, which is same as that of TDSM. The proposed query cut can also be used to split partitions using bottom- up (RB tree) techniques.

A. Top-Down Heuristic 1 (TDH1)

In TDSM, the partitions are split along the median. Consider a partition that overlaps a query. If the median also falls inside the query then even after splitting the partition, the imprecision for that query will not change as both the new partitions still overlap the query as illustrated. In this heuristic, we propose to split the partition along the query cut and then choose the dimension along which the imprecision is minimum for all queries [2]. If multiple queries overlap a partition, then the query to be used for the cut needs to be selected. The queries having imprecision greater than zero for the partition are sorted based on the imprecision bound and the query with minimum imprecision bound is selected. The intuition behind this decision is that the queries with smaller bounds have lower tolerance for error and such a partition split ensures the decrease in imprecision for the query with the smallest imprecision bound. If no feasible cut satisfying the privacy requirement is found, then the next query in the sorted list is used to check for partition split. If none of the queries allow partition split, then that partition is split along the median and the resulting partitions are added to the output after compaction. A feasible cut means that each partition resulting from split should satisfy the privacy requirement.

B. Top-Down Heuristic 2 (TDH2)

In the Top-Down Heuristic 2 algorithm, the query bounds are updated as the partitions are added to the

output. This update is carried out by subtracting the ic Qj Pi value from the imprecision bound BQj of each query, for a Partition, say Pi, that is being added to the output. For example, if a partition of size k has imprecision 5 and 10 for Queries Q1 and Q2 with imprecision bound 100 and 200, then the bounds are changed to 95 and 190, respectively. The best results are achieved if the k-d tree traversal is depth-first (preorder). Preorder traversal for the k-d tree ensures that a given partition is recursively split till the leaf node is reached. Then, the query bounds are updated. Initially, this approach favors queries with smaller bounds. As more partitions are added to the output, all the queries are treated fairly. During the query bound update, if the imprecision bound for any query gets violated, then that query is put on low priority by replacing the query bound by the query size. The intuition behind this decision is that whatever future partition splits TDH2 makes, the query bound for this query cannot be satisfied. Hence, the focus should be on the remaining queries.

C. Top-Down Heuristic 3 (TDH3)

The time complexity of the TDH2 algorithm is O(d 2 *Q* n 2), which is not scalable for large data sets (greater than 10 million tuples). In the Top- Down Heuristic 3 algorithm (TDH3, for short), we modify TDH2 so that the time complexity of O(d *Q n* lg *n*) can be achieved at the cost of reduced precision in the query results. Given a partition, TDH3 checks the query cuts only for the query having the lowest imprecision bound. Also, the second constraint is that the query cuts are feasible only in the case when the size ratio of the resulting partitions is not highly skewed. We use a skew ratio of 1:99 for TDH3 as a threshold. If a query cut results in one partition having a size greater than hundred times the other, then that cut is ignored.

VI. CONCLUSION

An accuracy-constrained privacy-preserving access control framework for relational data has been proposed. The framework is a combination of access control and privacy protection mechanisms. The access control mechanism allows only authorized query predicates on sensitive data. The privacy preserving module anonymizes the data to meet privacy requirements and imprecision constraints on predicates set by the access control mechanism. We formulate this interaction as the problem of k- anonymous Partitioning with Imprecision Bounds (k- PIB). We give hardness results for the k-PIB problem and present heuristics for partitioning the data to satisfy the privacy constraints and the imprecision bounds. In the current work, static access control and relational data model has been assumed. For future work, we plan to extend the proposed privacy- preserving access control to incremental data and cell level access control.

REFERENCES

[1] E. Bertino and R. Sandhu, "Database Security-Concepts, Approaches, and Challenges," IEEE Trans. Dependable and Secure Computing, vol. 2, no. 1, pp. 2-19, Jan.-Mar. 2005.

[2] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov. 2001.

[3] B. Fung, K. Wang, R. Chen, and P. Yu, "Privacy- Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, vol. 42, no. 4, article 14, 2010.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-Diversity: Privacy Beyond k-anonymity," ACM Trans. Knowledge Discovery from Data, vol. 1, no. 1, article 3, 2007.

[5] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization Techniques for Large-Scale Datasets," ACM Trans. Database Systems, vol. 33, no. 3, pp. 1-47, 2008.