International Journal of Science Engineering and Advance Technology

A Novel System for Privacy-Preserving Access Control for Relational Data with Accuracy

Sai Sirisha Chittineni[1], Bandarupalli Mouleswara Rao[2]
[1] M.Tech (CSE), NRI Institute of Technology, A.P., India.
[2]Assoc Professor, Dept. of Computer Science & Engineering, NRI Institute of Technology, A.P., India.

Abstract — Access Control is a set of controls to restrict access to certain resources. If we think about it, access controls are everywhere around us. A door to your room, the guards allowing you to enter the office building on seeing your access card, swiping your card and scanning your fingers on the biometric system, a queue for food at the canteen or entering your credentials to access FB, all are examples of various types of access control. Here we focus only on the logical Access Control mechanisms. Access control mechanisms protect sensitive information from unauthorized users. 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. A PPM can use suppression and generalization of relational data to anonymize and satisfy privacy requirements, e.g., k-anonymity and l-diversity, against identity and attribute disclosure. However, privacy is achieved at the cost of precision of authorized information. In this paper, we propose an accuracy-constrained privacy-preserving access control framework. 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. The techniques for workload-aware anonymization for selection predicates have been discussed in the literature. However, to the best of our knowledge, the problem of satisfying the accuracy constraints for multiple roles has not been studied before. In our formulation of the aforementioned problem, we propose heuristics for anonymization algorithms and show empirically that the proposed approach satisfies imprecision bounds for more permissions and has lower total imprecision than the current state of the art.

Keywords — Access control, privacy, k-anonymity, query evaluation.

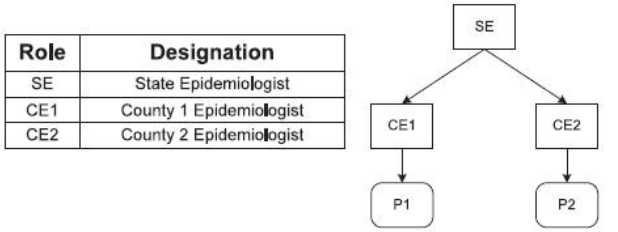
I. Introduction

Access Control Mechanisms: Discretionary Access Control (DAC) as the name suggests, this access control model is based on a user's discretion. i.e, the owner of the resource can give access rights on that resource to other users based on his discretion. Access Control Lists (ACLs)[2] are a typical example of DAC. Specifying the "rwx" permissions on a unix file owned by you is another example of DAC Most of the operating systems including windows, flavors of Unix are based on DAC[3] Model Mandatory Access Control (MAC): In this Model, users/owners do not enjoy the privilege of deciding who can access their files. Here the operating system is the decision maker overriding the user's wishes. In this model every Subject (users) and Object (resources) are classified and assigned with a security label. The security labels of the subject and the object along with the security policy determine if the subject can access the object. The rules for how subjects access objects are made by the security officer, configured by the administrator, enforced by the operating system, and supported by security technologies. This is a stricter and rather static Access Control model as compared to DAC and is mostly suited for military organizations where data classification and confidentiality is of prime importance. Special types of the UNIX operating systems are based on MAC model. Role Based Access Control (RBAC): RBAC is the buzzword across enterprises today. In this model the access to a resource is governed based on the role that the subject holds within an organization. RBAC [15] is also known as non-discretionary Access Control because the user inherits privileges that are tied to his role. The user does not have a control over the role that he will be assigned. Each of the above Access Models has its own advantages and disadvantages. The selection of the appropriate Access Model by an organization should be done by considering various factors such as type of business, no of users, organization's security policy etc. For implementing any access control, the two driving factors are 1. Least privilege principle i.e, the user should only have the minimum privileges to perform the tasks that he is supposed to do. 2. Segregation of duties i.e, having more than one user to perform a critical task so as to reduce the risk of internal frauds. As the no of users and the resources grow in an organization, it becomes extremely difficult to manage user's access rights through ACLs. It not only increases

the cost of administration but also results in granting of excess privileges to users thus violating the least privilege principle and hence exposing the organization to risks. Moreover, the complexity involved with this approach makes it too hard for any organization to comply with the regulatory compliances. So in organizations where the no of users and the employee turnover is large, RBAC [15] is the optimum solution for Access Control. By having privileges tied to roles, and users being assigned to these roles, makes it much simpler for an organization to manage the access to its resources. RBAC also fastens the employee on-boarding & de-boarding process by tying the provisioning/de-provisioning to the roles. Moreover, RBAC also makes the implementation of least privilege principle and SOD easier, hence helping the organizations in complying to the strict regulatory standards.

RBAC but the roles itself are vaguely defined? Or if proper consideration is not given to the privileges being assigned to Roles. The whole purpose of adopting RBAC stands defeated in this case. This is where Role Engineering comes into play. I thought before heading to role engineering it's important that the reader is aware of a few basic concepts related to access control. Organization collects and analyzes consumer data to improve their services. Access Control Mechanisms (ACM) are used to ensure that only authorized information is available to users. However, sensitive information can still be misused by authorized users to compromise the privacy of consumers. The concept of privacy-preservation for sensitive data can require the enforcement of privacy policies or the protection against identity disclosure by satisfying some privacy requirements [1]. 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 [5]. Anonymization algorithms use suppress on and generalization of rec rds to satisfy privacy requirements with minimal distortion of micro data. The anonymity techniques can be used with an access control mechanism to ensure both security and privacy of the sensitive information. The privacy is achieved at the cost of accuracy and imprecision is introduced in the authorized information under an access control policy. We use the concept of imprecision bound for each permission to define a threshold on the amount of imprecision that can be tolerated. Existing workloadaware anonymization techniques [5], [6] 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 (e.g., increasing the value of k or l) results in additional imprecision for queries. However, 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. 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. However, the concept of accuracy constraints for permissions can be applied to any privacy-preserving security policy, e.g., discretionary access control. Example1 (Motivating Scenario). Syndromic surveillance systems are used at the state and federal levels to detect and monitor threats to public health [7]. The department of health in a state collects the emergency department data

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



Permission	Authorized Query Predicate
P1	Location = County 1 ∧ Age = 15–65 ∧ Syndrome = Influenza
P2	Location = County 2 ∧ Age = 15–65 ∧ Syndrome = Influenza

Figure 1 Access Control Policy

(age, gender, location, time of arrival, symptoms, etc.) from county hospitals daily. Generally, 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. An access control policy is given in Fig. 1 that allows the roles to access the tuples under the authorized predicate, e.g., Role CE1 can access tuples under Permission P1. The epidemiologists at the state and county level suggest community containment measures, e.g., isolation or quarantine according to the number of persons infected in case of a flu outbreak. According to the population density in a county, an epidemiologist can advise isolation if the number of persons reported with influenza are greater than 1,000 and quarantine if that number is greater than 3,000 in a single day. The anonymization adds imprecision to the query results and the imprecision bound for each query ensures that the results are within the tolerance required. If the imprecision bounds are not satisfied then unnecessary

false alarms are generated due to the high rate of false positives.

First, we formulate the accuracy and privacy constraints as the problem of k-anonymous Partitioning with Imprecision Bounds (k-PIB) and give hardness results. Second, we introduce the concept of accuracy-constrained privacy-preserving access control for relational data. Third, we propose heuristics to approximate the solution of the k-PIB problem and conduct empirical evaluation.

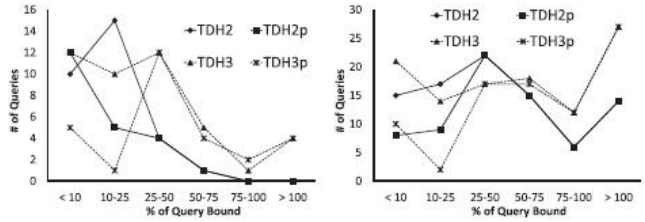
Given a relation T= {A1,A2, . . .,An}, where Ai is an attribute, T* is the anonymized version of the relation T. We assume that T is a static relational table. The attributes can be of the following types: Identifier. Attributes, e.g., name and social security that can uniquely identify an individual. These attributes are completely removed from the anonymized relation. Quasi-identifier(QI). Attributes, e.g., gender, zip code, birth date that can potentially identify an individual based on other information available to an adversary. QI attributes are generalized to satisfy the anonymity requirements. Sensitive attribute. Attributes, e.g., disease or salary, that if associated to a unique individual will cause a privacy breach.

II .PROBLEM STATEMENT

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. A PPM can use suppression and generalization of relational data to anonymize and satisfy privacy requirements, e.g., k-anonymity and l-diversity, against identity and attribute disclosure. However, privacy is achieved at the cost of precision of authorized information.

III .RELATED WORK

Access control mechanisms for databases allow queries only on the authorized part of the database [8], [10]. Predicate based fine-grained access control has further been proposed, where user authorization is limited to pre-defined predicates [11]. Enforcement of access control and privacy policies has been studied in [2]. 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 [14]. They use the definition of differential privacy [15] 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. [16] 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



(a) $k = 5$ and Bound = 30% (b) $k = 7$ and Bound = 30%

Figure 3 Improvements after repartitioning for k-anonymity for the Census data set.

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. [14]. They have proposed the Selection Mondrian algorithm, which is a modification to the greedy multidimensional partitioning algorithm Mondrian [13]. 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 Rþ-tree based anonymization algorithm [6]. The authors illustrate by experiments that anonymized data using biased Rþ-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 [7]. They also introduce the problem of accuracy-constrained anonymization for a given bound of acceptable information loss for each equivalence class [8]. Similarly, Xiao et al. [1] 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.

Access Control for Relational Data

Fine-grained access control for relational data allows to define tuple-level permissions, e.g., Oracle VPD [8] and SQL [9]. For evaluating user queries, most approaches assume a Truman model [10]. In this model, a user query is modified by the access control mechanism and only the authorized tuples are returned. Column level access control allows queries to execute on the authorized column of the relational data only [8], [11]. Cell level access control for relational data is implemented by replacing the unauthorized cell values by NULL values [12]. Role-based Access Control (RBAC) allows defining permissions on objects based on roles in an organization. An RBAC policy configuration is composed of a set of Users (U), a set of Roles (R), and a set of Permissions (P). For the relational RBAC model, we assume that the selection predicates on the QI attributes define permission [11]. UA is a user-to-role (U R) assignment relation and PA is a roleto-permission (R P) assignment relation. A role hierarchy (RH) defines an inheritance relationship among roles and is a partial order on roles (R R) [13].Each permission defines a hyper-rectangle in the tuple space and all the tuples enclosed by this hyper-rectangle are authorized to the role assigned to the permission. In practice, when a user assigned to a role executes a query, the tuples satisfying the conjunction of the query predicate and the permission are returned [1], [10].

Definition 1 (Equivalence Class (EC)). An equivalence class is a set of tuples having the same QI attribute values.

Definition 2 (k-anonymity Property). A table T^* satisfies the
k-anonymity property if each equivalence class has k or more
tuples [2].

k-anonymity is prone to homogeneity attacks when the sensitive value for all the tuples in an equivalence class is the same. To counter this shortcoming, l-diversity has been proposed [4] and requires that each equivalence class of

	QI_1	QI_2	S_1
ID	Age	Zip	Disease
1	5	15	Flu
2	15	25	Fever
3	28	28	Diarrhea
4	25	15	Fever
5	22	28	Flu
6	32	35	Fever
7	38	32	Flu
8	35	25	Diarrhea

(a) Sensitive table

	QI_1	QI_2	S_1
ID	Age	Zip	Disease
1	0-20	10-30	Flu
2	0-20	10-30	Fever
3	20-30	10-30	Diarrhea
4	20-30	10-30	Fever
5	20-30	10-30	Flu
6	30-40	20-40	Fever
7	30-40	20-40	Flu
8	30-40	20-40	Diarrhea

(b) 2-anonymous Table

Figure 2 Generalization for K-anonymity and l-diversity

T^* contain at least l distinct values of the sensitive attribute. For sensitive numeric attributes, an l-diverse equivalence class can still leak information if the numeric values are close to each other. For such cases, variance diversity [5] has been proposed that requires the variance of each equivalence class to be greater than a given variance diversity parameter. The table in Fig. 2a does not satisfy k-anonymity because knowing the age and zip code of a person allows associating a disease to that person. The table in Fig. 2b is a 2-anonymous and 2-diverse version of table in Fig. 2a. The ID attribute is removed in the anonymized table and is shown only for identification of tuples. Here, for any combination of selection predicates on the zip code and age attributes, there are at least two tuples in each equivalence class. In Section 4, algorithms are presented for kanonymity only. However, the experiments are performed for both l-diversity and variance diversity using the proposed heuristics for partitioning.

IV. Conclusion

The access control mechanism allows only authorized query predicates on sensitive data. The privacypreserving 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 problemof k-anonymous Partitioning with Imprecision Bounds (k-PIB).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.. We give hardness results for the k-PIB problem and present heuristics for partitioning the data to the 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.
[6] T. Iwuchukwu and J. Naughton, "K-Anonymization as Spatial Indexing: Toward Scalable and Incremental Anonymization," Proc. 33rd Int'l Conf. Very Large Data Bases, pp. 746-757, 2007.
[7] J. Buehler, A. Sonricker, M. Paladini, P. Soper, and F. Mostashari, "Syndromic Surveillance Practice in the United States: Findings from a Survey of State, Territorial, and Selected Local Health Departments," Advances in Disease Surveillance, vol. 6, no. 3, pp. 1-20, 2008.
[8] K. Browder and M. Davidson, "The Virtual Private Database in oracle9ir2," Oracle TechnicalWhite Paper, vol. 500, 2002.
[9] A. Rask, D. Rubin, and B. Neumann, "Implementing Row-and Cell-Level Security in Classified Databases Using SQL Server
2005," MS SQL Server Technical Center, 2005.
[10] S. Rizvi, A. Mendelzon, S. Sudarshan, and P. Roy, "Extending Query Rewriting Techniques for Fine-Grained Access Control," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 551-562,
2004.
[11] S. Chaudhuri, T. Dutta, and S. Sudarshan, "Fine Grained Authorization through Predicated Grants," Proc. IEEE 23rd Int'l Conf. Data Eng., pp. 1174-1183, 2007.
[12] K. LeFevre, R. Agrawal, 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.
[13] D. Ferraiolo, R. Sandhu, S. Gavrila, D. Kuhn, and R. Chandramouli, "Proposed NIST Standard for Role-Based Access Control," ACM Trans. Information and System Security, vol. 4, no. 3, pp. 224- 274, 2001.
[14] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," Proc. 22nd Int'l Conf. Data Eng., pp. 25- 25, 2006.
[15] J. Friedman, J. Bentley, and R. Finkel, "An Algorithm for Finding Best Matches in Logarithmic Expected Time," ACM Trans. Mathematical Software, vol. 3, no. 3, pp. 209-226, 1977.
[16] A. Meyerson and R. Williams, "On The Complexity of Optimal
k-Anonymity," Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems, pp. 223-228, 2004.