# Access Control Mechanism Using k-Anonymity Algorithm for Relational Data

[1]Yadali Bhavana,[2]S Phani Praveen,[3]N V Ramana Gupta
[1]Student,[2]Asst. Professor,[3]Asst. Professor
[1] Department of Computer Science and Engineering,
[1]Prasad V Potluri Siddhartha Institute of Technology, Kanuru, India

***Abstract* -- Information security is the practice to protect the confidentiality, integrity, and availability of computer system data from those with malicious intentions. Privacy-preserving in data mining using access control mechanism which is a specialized technique used for providing the privacy of the sensitive data in such a way that only authorized user can access his own information if the user tries to access any unauthorized information it will appear in an anonymous format. This type of technique in anonymizing the data is mainly used in security management system where privacy is to be maintained regarding the details of the user because of sensitivity content in the information. So the proposed application uses certain mechanisms to protect the privacy and integrity of data. There are two types of approaches: firstly Suppression and Generalization for anonymizing the data and Access Control Mechanism (ACM) for preventing unauthorized access of information and hence maintaining Privacy by preventing identity disclosure of the person. Here techniques such as k-anonymity and l-diversity are used against identity and attribute disclosure such as Privacy Preserving Mechanism (PPM) for protecting the privacy requirements. The anonymity 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 under an access control policy; imprecision is introduced in the authorized information.**

***Index Terms* -- Information security, k-anonymity, l-diversity, Suppression, Generalization.**

## I. INTRODUCTION

To improve the quality of services, Organizations collect the user data and analyze the information. Access Control Mechanism (ACM) is used to guarantee that only approved information is available to consumers. On the other hand, approved users will still use the sensitive records of the customers to negotiate the privacy. The proposal of privacy-preservation designed for sensitive data will need the imposition of security against identity exposure or the privacy policies by fulfilling a few privacy necessities. The sensitive records still remain vulnerable to some of the attacks (i.e., linking attacks) by the approved user even though the distinctive attributes are removed. This dilemma has been studied extensively in the part of privacy definitions and micro-data disclosing, e.g., l-diversity, k-anonymity, and variance diversity. Here, we inspect privacy-preservation of the user as of anonymity prospect. Anonymization algorithms apply generalization and suppression on user data records to assure privacy needs by lowest distortion of micro-data.

Various organizations are publishing the micro-data for several different purposes; for example, public health research, demographic research, business, etc. Hence, the privacy of an individual is at risk due to this type of published data. Data holders remove or encrypt the precise identifiers like phone numbers, names, addresses, and social security number, in order to preserve the anonymity of data entities. Other attributes like date of birth, race, sex, zip code, etc are yet used to recognize the anonymous individuals when shared together with visibly released information. The large quantity of information that is easily available today can be a serious problem if the attackers use the improved computations.

To ensure the privacy and security of user's sensitive information, mechanisms like Access Control Mechanisms are used by the anonymity techniques. Imprecision is introduced in approved information under an access control policy. The privacy is achieved at the cost of accuracy.

## II. RELATED WORK

In the proposed approach, the algorithm uses the theory of imprecision bound for every permission to describe the threshold on the quantity of imprecision it can tolerate. Existing workload-aware anonymization techniques minimize the imprecision aggregate for all queries and the imprecision added to each query in the anonymized micro data is not known. Additional imprecision for queries make the privacy requirement much stricter.

Earlier, it is not studied that the problem of fulfilling accuracy constraints for individual queries in a policy. The anonymization for continuous data publishing has been studied in the literature. The heuristics projected for access control mechanism for relational data are also relevant in the context of workload-aware anonymization. In the proposed system, anonymizing the static relational table at once is the focal point. Role-based access control is assumed to illustrate the approach. However, the notion of accuracy constraints for permissions can be applied to any privacy-preserving security policy, e.g., discretionary access control.

Firstly, we formulate the privacy and accuracy constraints as the problem of k-anonymous Partitioning with Imprecision Bounds (k-PIB) also hardness results are given. After that, the concept of access control mechanism using k-anonymity algorithm for relational data is introduced. Finally, heuristics to approximate the solution of the k-PIB is proposed and empirical evaluation is conducted.

## III. ACCESS CONTROL MECHANISM FOR RELATIONAL DATA

An accuracy-constrained privacy-preserving access control mechanism [1] shown in Fig:1 is proposed. The privacy protection mechanism ensures that the privacy and the accuracy objectives are met before the sensitive data is available to the access control mechanism. The queries in the access control workload are based on selection predicates on the Quasi Identifier attributes. Permissions and imprecision bounds for every permission, role to- permission assignments, and user-to-role assignments are defined by the administrator.

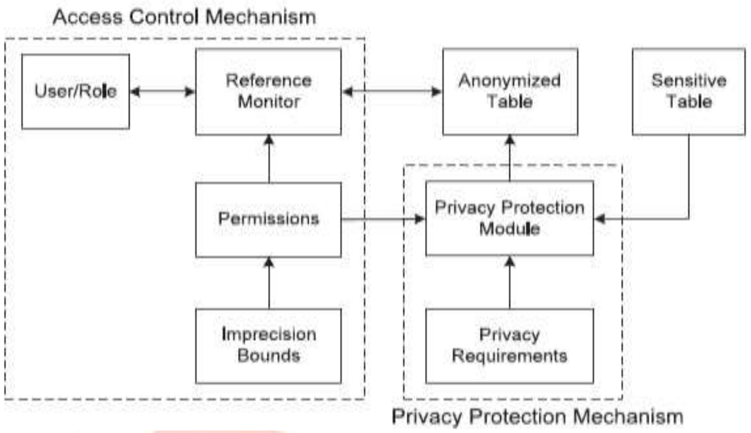


Fig:1. Accuracy-constrained privacy preserving access control mechanism

The specification of the imprecision bound ensures that the approved data has the desired level of accuracy. Approved query predicates are allowed by Access Control Mechanism for sensitive data. The imprecision bound for every query along with privacy requirement should be met by privacy protection mechanism. Information about imprecision bound is kept private to avoid the privacy issues.

Access Control Mechanism protects sensitive data from unauthorized users. Selection predicates are defined by access control approaches that are accessible to roles. Role-based Access Control (RBAC) is used to define the permissions on object based on roles in an organization. In Cell level access control, user who satisfies the access policy and having valid set of attributes can access the data. Admin has authority to allow attributes. [8].

- **User/Role:** It allows defining permissions on objects based on roles in an organization. An RBAC is composed of a set of Roles, a set of Users and a set of Permissions.
- **Permissions:** The imprecision bound for user-to- role assignments, each query and role-to permission assignments. It based on selection predicates on the Quasi Identifier attributes.
- **Imprecision Bound:** The imprecision bound can be used to meet the privacy requirement. It ensures that the authorized data has the desired level of accuracy.

### 3.1. ANONYMITY DEFINITIONS:

- **Equivalence Class:** An equivalence class is a set of tuples having the same Quasi Identifier attribute values.
- **k-anonymity Property:** A table T* satisfies the k-anonymity property if each equivalence class has k or more tuples [1].

If the sensitive value is same for all the tuples in an equivalence class, then homogeneity attacks suffers k- Anonymity. The l-diversity has been proposed to overcome the disadvantage of k-anonymity. It requires that every equivalence class of T* should have minimum 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. Variance diversity has been proposed for such cases that requires the variance of each equivalence class to be greater than a given variance diversity parameter.

|  | $QI_1$ | $QI_2$ | $S_1$ |
|---|---|---|---|
| ID | AGE | ZIP | DISEASE |
| 1 | 5 | 15 | Flue |
| 2 | 15 | 25 | Fever |
| 3 | 28 | 28 | Diarrhea |
| 4 | 25 | 15 | Fever |
| 5 | 22 | 28 | Flue |
| 6 | 32 | 25 | Fever |
| 7 | 38 | 32 | Flue |
| 8 | 35 | 25 | Diarrhea |

(a)

|  | $QI_1$ | $QI_2$ | $S_1$ |
|---|---|---|---|
| ID | AGE | ZIP | DISEASE |
| 1 | 0-20 | 10-30 | Flue |
| 2 | 0-20 | 10-30 | Fever |
| 3 | 20-30 | 10-30 | Diarrhea |
| 4 | 20-30 | 10-30 | Fever |
| 5 | 20-30 | 10-30 | Flue |
| 6 | 30-40 | 20-40 | Fever |
| 7 | 30-40 | 20-40 | Flue |
| 8 | 30-40 | 20-40 | Diarrhea |

(b)

Fig:2. Generalization for k-anonymity and l-diversity

The table in Figure 2(a) 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 Figure 2(b) is a 2-anonymous and 2-diverse version of table in Figure 2(a). The ID attribute is encrypted in the anonymized table and is shown only for identification of tuples. Here, for any sequence of selection predicates on the zip code and age attributes, there are at least two tuples in each equivalence class.

- **Explicit Identifier:** Attributes, e.g., name and social security number that can uniquely identify an individual [4].These attributes are encrypted by using Rijndael symmetric encryption algorithm from the anonymized relation.
- **Quasi-Identifier:** QI attributes are generalized to satisfy the anonymity requirements. Attributes, e.g., gender, zip code, birth date, which can identify an individual based on other information available to an adversary.
- **Sensitive Attribute:** Attributes, e.g., disease or salary, that if associated to a unique individual will cause privacy break.

### 3.2. Predicate Evaluation and Imprecision:

For query predicate evaluation over a table, say T, a tuple is included in the result if all the attribute values satisfy the query predicate. Here, we only consider conjunctive queries, where each query can be expressed as a d-dimensional hyper-rectangle. The semantics for query evaluation on an anonymized table T* needs to be defined. When the equivalence class partition is fully enclosed inside the query region, all tuples in the equivalence class are part of the query result.

- **Query Imprecision:** 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 is known as Query Imprecision [5]. Imprecision for query $Q_i$ is denoted as imp $Q_i$.

### 3.3 Anonymization with Imprecision Bounds:

- **Query Imprecision Bound:** Total imprecision acceptable for a query predicate $Q_i$ and is preset by the access control administrator and is denoted by $BQ_i$.

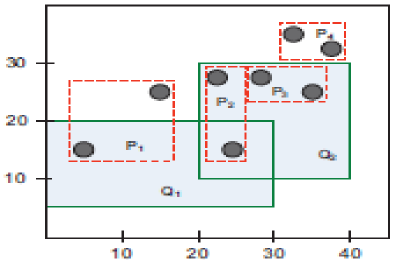


Fig:3. Anonymization satisfying imprecision bounds

## IV. TOP-DOWN HEURISTIC ALGORITHM

**Step-1:** Initialize the set of candidate partition.

**Step-2:** Sort the queries in which candidate partitions are overlapping with imprecision higher than zero.

**Step-3:** Select the least imprecision bound queries.

**Step-4:** Checks for the possible split of the partition along the query interval.

**Step-5:** The resultant partitions are added to the candidate partition when a possible cut is found.

**Step-6:** The candidate partition is checked for the median cut when possible cut is not found.

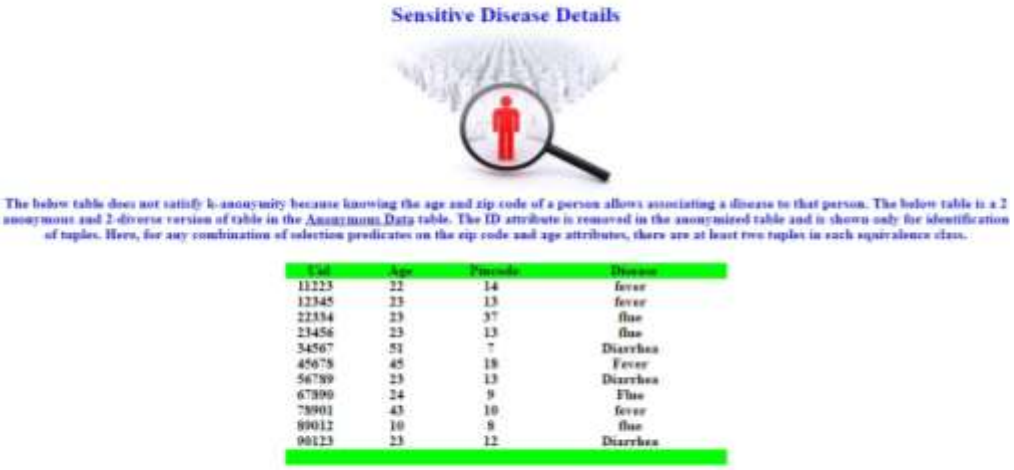
The objective of Top-Down Heuristic Algorithm is to minimize the total imprecision for all queries while the imprecision bounds for queries have not been considered. Initially, the whole tuple space is taken as one partition and then partitions are divided recursively until the time new partitions meet the privacy requirement. To divide a partition, two decisions need to be made, i) Choosing a split value along each dimension, and ii) Choosing a dimension along which to split. The dimension is selected along which the sum of imprecision for all queries is minimum. The split value is chosen along the median.

The heuristic algorithm will helps to provide the secured access control mechanism. The imprecision bound is set by the admin and is not known to the user. So it provides the secured access control method.

## V. RESULT ANALYSIS

In this approach, along the query cut, the partition splitting is proposed and then dimension is chosen for all queries in which the imprecision is least. If multiple queries overlap a partition, then the query for the cut needs to be selected. On the basis of imprecision bound, if the queries having imprecision more than zero for the partition are sorted and then the query with minimum imprecision bound is selected.

The reason 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 none of feasible cut satisfying the privacy requirement is found, then the next query in the sorted list is used to check for partition split. Partition will split along the median and the resulting partitions are added to the output after compaction in the case where none of the queries allow partition split.

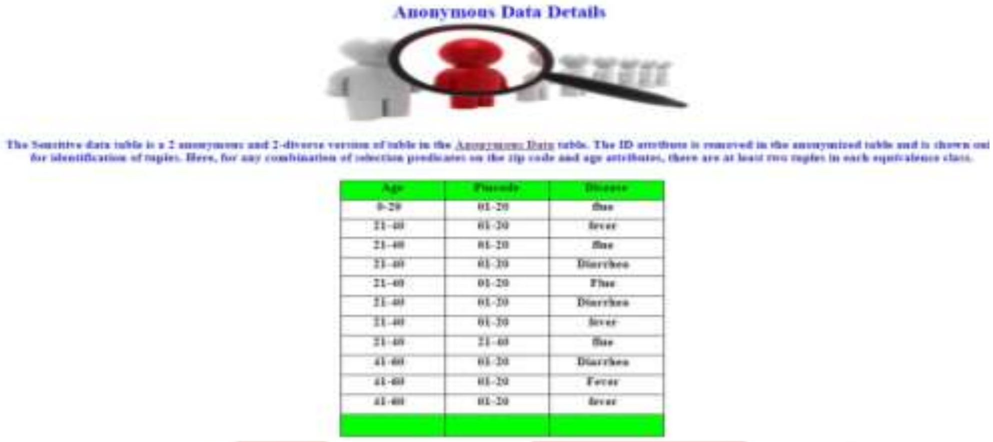


Sensitive Disease Details

The below table does not satisfy k-anonymity because knowing the age and zip code of a person allows associating a disease to that person. The below table is a 2 anonymous and 2-diverse version of table in the Anonymous Data table. 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.

| Uid | Age | Pincode | Disease |
|---|---|---|---|
| 11223 | 22 | 14 | fever |
| 12345 | 23 | 13 | fever |
| 22334 | 23 | 37 | flue |
| 23456 | 23 | 13 | flue |
| 34567 | 51 | 7 | Diarrhea |
| 45678 | 45 | 19 | Fever |
| 56789 | 23 | 13 | Diarrhea |
| 67890 | 24 | 9 | Flue |
| 78901 | 43 | 10 | fever |
| 89012 | 10 | 8 | flue |
| 90123 | 23 | 12 | Diarrhea |

Fig:4. Sensitive Data



Anonymous Data Details

The Sensitive data table is a 2 anonymous and 2-diverse version of table in the Anonymous Data table. The ID attribute is removed in the anonymized table and is shown onl 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.

| Age | Pincode | Disease |
|---|---|---|
| 0-20 | 01-20 | flue |
| 21-40 | 01-20 | fever |
| 21-40 | 01-20 | flue |
| 21-40 | 01-20 | Diarrhea |
| 21-40 | 01-20 | Flue |
| 21-40 | 01-20 | Diarrhea |
| 21-40 | 01-20 | fever |
| 21-40 | 21-40 | flue |
| 41-60 | 01-20 | Diarrhea |
| 41-60 | 01-20 | Fever |
| 41-60 | 01-20 | fever |

Fig:5. Anonymous Data

The proposed system gets information in an anonymous version of sensitive table. In the anonymized table, ID attribute is removed. There are minimum two tuples in every equivalence class for any set of selection predicates on the zip code and age attributes.

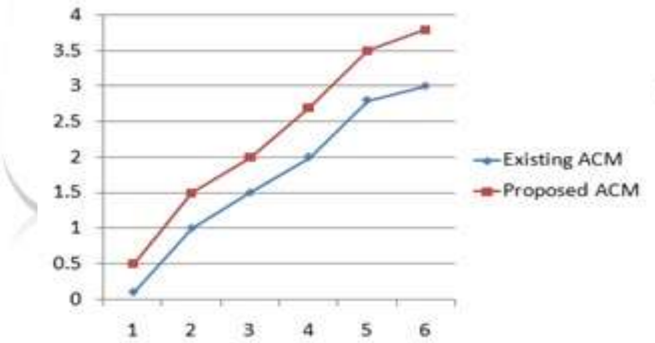


Fig:6. Graph between Proposed ACM Existing ACM approaches

## VI. CONCLUSION

An access control mechanism using k-anonymity algorithm for relational data has been proposed. Access control mechanism with privacy protection mechanism is defined in the framework. The access control mechanism grants only authorized query predicates on sensitive data. The privacy-preserving mechanism anonymizes the data to meet requisites of privacy, and also Access control mechanism are setup on imprecision constraints on predicates set. The hardness solution for the k-PIB problem and the heuristics for partitioning the data to satisfy the privacy constraints and imprecision bounds are given. In the proposed work, static access control and relational data model has been assumed.

## REFERENCES

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

[2] ZahidPervaiz, Walid G.Aref, Arif Ghafoor, and Nagabhushana Prabhu "Accuracy-Constrained Privacy Preserving Access Control Mechanism for Relational Data" IEEE Trans. On Knowledge and Data Engineering, Vol. 26, No. 4, April 2014.

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

[4] S. Chaudhuri, R. Kaushik, and R. Ramamurthy, “Database Access Control & Privacy: Is There a Common Ground?” Proc. Fifth Biennial Conf. Innovative Data Systems Research (CIDR), pp. 96-103, 2011.
[5] K. LeFevre, D. DeWitt, and R. Ramakrishnan,“Workload-Aware Anonymization Techniques for LargeScale Datasets,” ACM Trans. Database Systems, vol. 33, no. 3, pp. 1-47, 2008.
[6] D. Ferraiolo, R. Sandhu, S. Gavrila, D. Kuhn, and R. Chandramouli, “Proposed NIST Standard for RoleBased Access Control,” ACM Trans. Information and System Security, vol. 4, no. 3, pp. 224- 274, 2001.
[7] 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.
[8] Li, N., Li. T. and Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In 23rd International Conference on Data Engineering. pp. 106-115. IEEE Computer Society (2007).