ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 3; Issue 4; July-August-2015; Page No. 74-81

Privacy-Preserving Access Control Mechanism for Social Statistical Data considering Accuracy Constrained

Kaushlendra Tiwari[1], Awadhesh Kumar Sharma[2]

[1]Department of Computer Science & Engineering, MMMUT, Gorakhpur

Kaushlendra38@gmail.com

[2] Associate Professor, Department of Computer Science & Engineering, MMMUT, Gorakhpur

akscse@rediff.com

Abstract

In the current era of digitization, data volume is increasing rapidly. In order to store and manage these data for different purposes, there exist various standards and policies. Since data storing and managing is a tedious task, some organization may adopt data outsourcing to third party in order to reduce their load and to concentrate on quality of services. It becomes a privacy breach of users when third party cannot be trusted. In order to ensure privacy of users, different encryption approaches are used such as k-anonymity, l-diversity, variance diversity etc. There are two basic approaches to ensure the privacy of users. First is encrypt the whole database before storing at third party side and whenever needed import, decrypt and use them. Second approach is data fragmentation, in which whole data is divided in to n blocks and stored in different databases. We can use different fragments of different database at same time in order to process a query. Both approaches have their own pros and cons. Besides this, access control policy also plays vital role in order to ensure the privacy of user. We can achieve privacy using k-anonymity, l-diversity or variance diversity at the cost of data imprecision. Higher the level of imprecision, lower the accuracy will be. In this paper we are concerning about user privacy especially data stored on social media sites like facebook, twitter and others. Personal information of users is still susceptible to privacy breach, although they have their own privacy protection mechanism and access control policies with limitations of their own. In this paper we are providing privacy preserving access control mechanism for social statistical data considering accuracy constrained. It ensures the privacy of user data and provides data imprecision with more accuracy.

Keywords- Access control policy, Privacy protection mechanism, k-anonymity, l-diversity

I. INTRODUCTION

When sensitive information is shared, the privacy of this information may be breach by a person which is authorized or unauthorized. There are various access control and privacy protection policies which prevent the unauthorized users to access the data, but an authorized user still susceptible to privacy breach. Privacy Protection Mechanism (PPM) can satisfy privacy requirements such as k-anonymity and l-diversity with its suppression and generalization of relational data. While satisfying the privacy requirement, k-anonymity [3] [11] or l-diversity [10], selection predicates are defined by the access control policies which is available to rolls [16]. Privacy Protection Mechanism (PPM) should also satisfy the imprecision bounds for each selection predicate. However, the information will become precise and hence increases inaccuracy. Thus access control mechanism [4] protects the sensitive information from unauthorized user, but Privacy Protection Mechanism (PEM) protects the privacy of users from both unauthorized and authorized user [2].

There are two basic approaches [1] to ensure the privacy [5] of users. First is encrypt the whole database before storing at third party side and whenever needed import, decrypt and use them. Second approach is data fragmentation, in which whole data is divided in to *n* blocks and stored in different databases. We can use different fragments of different database at same time in order to process a query. Both approaches have their own pros and cons. Besides this, access control policy [6], [12] also plays vital role in order to ensure the privacy of user. In this paper we are concerning about user privacy especially data stored on social media sites like facebook, twitter and others. Personal information of users is still susceptible to privacy breach, although they have their own privacy and access control policies with limitations given above. In this paper we are providing privacy preserving access control mechanism for social statistical data considering accuracy

Corresponding author: Kaushlendra Tiwari

constrained. It ensures the privacy of user data and provides data imprecision with more accuracy.

1.1. Terms and definitions

Given a relation R = {A1; A2; . . .; An}, where Ai is an attribute, A* is the anonymized version of the relation R. Here, we have considered a relational table R which is a statically maintained [2]. The table attributes may be of different types as given below:

(a) Identifier attributes- e.g., name and social security, which can identify a person uniquely. These attributes are not considered in anonymized relation.

(b) Quasi identifier (QI) attributes- e.g., zip code, Date of Birth, gender, etc. This can potentially identify a person based on some other information available to an opponent. Generalization of QI attributes satisfies the anonymity requirements.

(c) Sensitive attributes- e.g. salary, disease etc., which will cause privacy breach if it is belonging to a unique person.

1.2. K-anonymity-

In order to avoid the adversary to directly observe values in the database, we can add some noise into values in the database. In other words, when user queries the database, the database does not answer the correct result to the user. The results are returned to make some modification in order to disguise the correct data. Let V={v_1, v_2, ..., v_n } be a subset of values in attributes *A*, $V \subseteq A$, and values in *V* will mapping to *V'* . Multiple values correspond to the same value is generalized. Hence, adversary only sees value *V* in attributes, rather than *V*1, *V*2,...., *Vn* [1].

ID	Age	Work Class	Education
1	11	State-gov	Bachelors
2	18	Self-emp	Bachelors
3	25	Self-emp	Masters
4	35	Private	Masters
5	38	Federal-gov	Bachelors

(a) Table with sensitive data

ID	Age	Work Class	Education
1	10-20	State-gov	Bachelors
2	10-20	Self-emp	Bachelors
3	20-30	Self-emp	Masters
4	30-40	Private	Masters
5	30-40	Federal-gov	Bachelors

(b) Anonymize data on attribute Age

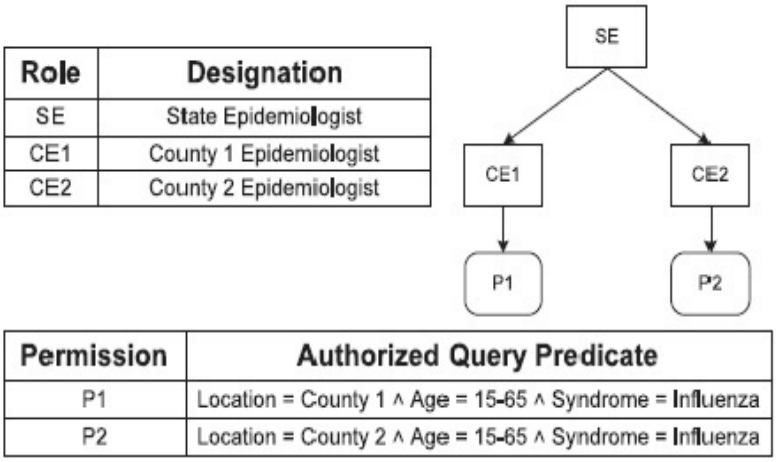
Fig. 1.2.1: Anonymization process

K-Anonymity [9] means if the information for each person contained in the data release, then it cannot be recognized by at least (k-1) individuals whose information also appears in the same data release. Consider the above table which shows the anonymization of data considering the attribute 'Age' as Quasi Identifier (QI) attribute. QI attribute can be selected according to the given database.

II. PROPOSED WORK

In this paper we are concerning about user privacy especially data stored on social media sites like facebook, twitter and others. Personal information of users is still susceptible to privacy breach, although they have their own privacy and access control policies with their own limitations. There are many approaches available for access control and privacy preserving of database. Here we are providing access control policy to access the database and implementing a privacy protection mechanism for social statistical dataset used to store sensitive information of a set of persons. In this paper we are providing privacy preserving access control mechanism for social statistical data considering accuracy constrained. It ensures the privacy of user data and provides data imprecision with more accuracy. For this purpose, we are using Top-Down Heuristic algorithms to anonymize data and enforcing accuracy constrained. These algorithms have never been implemented before for this purpose. The whole paper is divided in five section including Introduction, Background, Algorithm implementation and result, and finally Conclusion section.

Motivating Scenario- Consider the Syndromic surveillance systems which are used at regional or state levels to detect and monitor the threats which are vulnerable to public health [8]. The health department in a state collects the data related to the patients from emergency department like location, time of arrival, age, gender, symptoms, etc. from regional hospitals daily. Generally, daily updates consist of a static instance which falls under the category of syndrome as recognised by the health department. After that, anonymization of surveillance data is done in order to preserve the privacy of patients before sharing the data with departments of health at each region. An access control policy is given in Fig. 1. It allows only authorized access to the roles. Here, Role CE1 can access tuples under the permission P1.



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

Fig. 2.1: Access control policy

The epidemiologists at the state or regional level suggest community to take preventive or curative measures, e.g., isolation with infected person/area or quarantine according to the number of persons infected. According to the population density in a region, an epidemiologist considers the population of that area and advises the isolation, if the number of persons reported with influenza is greater than certain defined level, e.g., 1000, or suggest for quarantine if that number is greater than 3,000 in a single day. The process of anonymizing the data adds imprecision to the query results i.e. inaccuracy and for each query, the imprecision bound ensures that the results are tolerable. If the imprecision bounds are not satisfied then there may be generation of false alarms due to the high rate of false positives.

Here, disease related to a person is sensitive attribute when there is a restriction to reveal the patient name with associated disease due to some medical reasons. Similarly, there are also some data on social media for which a user does not want to reveal publically since it is related to their privacy. For this reason, there must be a Privacy Protection Mechanism to ensure the privacy of users.

III. BACKGROUND

Database must have some access control policy for defining the access level of users. It also must have some privacy policy related to data stored on database**.**

3.1. Access control policy

Only authorized query predicates are allowed by access control mechanism on sensitive data. There may be different level of access control exist for same database whether it can be strict or relaxed. Access control for relational data at fine grained level allows defining tuple-level access, e.g., Oracle VPD and SQL. In order to evaluate the user queries, most approaches follow a Truman model [15]. This model allows modifying the user query by the access control mechanism and returns only the authorized tuples. Access control may be implemented at Column level, in which queries are executed on authorized column of the relational data only or there may be an access control at cell level [13]. Role-based Access Control (RBAC) [16] defines permissions on objects considering roles available in an organization. A Roll Based Access Control policy configuration is combination of a set of Users (U), a set of Roles (R), and a set of Permissions (P) [16], [18]. For the relational Role-based Access Control (RBAC), we assume that the selection predicates on the attributes define permission.

We are using the concept of imprecision bound for each permission which defines a threshold on the amount of imprecision. The amount of imprecision to the provided data must be tolerable. Currently existing workload aware anonymization techniques [7] for all queries minimizes the imprecision aggregate and the imprecision added to each permission in the anonymized micro data [17] is not known.

3.2. Anonymization and accuracy

k-anonymization each record in attributes is identical with at least (*k*-1) other records. For this reason, adversary cannot identify each entity through observation directly. Since the level of data anonymization is depending on the value of 'k', the selection of this value is very important which should be high for high privacy. But on the same way, we are losing accuracy of data also. There should be a trade-off between privacy and data accuracy while selecting the value of 'k'.

3.3. Accuracy constrained access control mechanism

Accuracy constrained access control mechanism [2] given in figure 3.2.1.

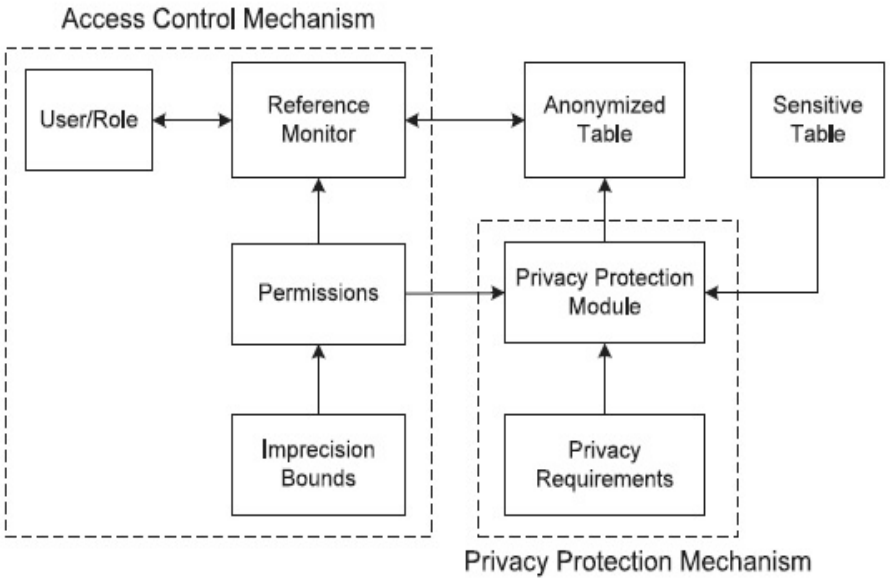


Fig. 3.2.1: Accuracy constrained access control mechanism

It is the combination of two separate modules (dataflow given by arrow) mentioned as access control mechanism and privacy protection mechanism. The first one manages the permissions to access the database and the second one ensures the privacy and accuracy goals before the data is available to the access control mechanism. Access control policy defines the permissions based on selection predicates on the QI attributes. It is the task of policy administrator to define the permissions considering appropriate imprecision bound for each permission/query, user-to-role assignments, and role-to-permission assignments [16]. The level of access control may be strict or relaxed according to the sensitivity of data [2]. The imprecision bound specification ensures the authorized data has the desired level of accuracy. The information of imprecision bound is not shared with the users because; if user knows the imprecision bound then it can result in violating the privacy requirement. There is a requirement of privacy protection mechanism in order to meet the privacy requirements along with the imprecision bound for each permission.

3.4. The k-PIB Problem

In k-anonymization, the value of k is hard to be determined as it shows the trade-off between degree of suppression and the accuracy. The optimal k-anonymity problem is a NP-complete problem for generalization and suppression [14]. The hardness result for k-anonymous Partitioning with Imprecision Bounds follows the construction of LeFevre et al. [11] that describes the hardness of k-PIB partitioning with the smallest average equivalence class size. For minimum number of queries, finding k-anonymous partitioning that violates imprecision bounds is also considered as a NP-hard problem [2].

Given a multiset of tuples, this is transformed into an equivalent set of distinct pairs (tuple; count). The cardinality of Query Q_i is the sum of count values of tuples falling inside the query hyper-rectangle. The constant q_v defines an upper bound for the number of queries that can violate the bounds. The decisional k-PIB problem is given as follows:

Decisional k-anonymity with Imprecision Bounds- Consider a set s ϵ S of unique pairs (tuple, count). here tuples are in the d-dimensional space and a set of queries Qi ϵ Q with imprecision bounds BQi , does there exist a multidimensional partitioning for S such that the size of every multidimensional partition Ri is greater than or equal to k and the number of queries violating imprecision bounds is less than the positive constant qv? [1], [11].

3.5. Heuristics for data partitioning

Pervaiz et. al. Provided top-down heuristics (TDH) [2] for accuracy constrained privacy preserving access for relational data. They provided three algorithms as TDH1, TDH2, and TDH3. They provided a framework for accuracy constrained privacy preserving access for relational database and here we are using it to implement on social statistical database which is especially for analysis about the social status of persons residing in certain area.

IV. ALGORITHMS IMPLEMENTATION AND RESULT

We are using top down heuristics to anonymize the given set of data. Zahid Pervaiz et. al. developed top down heuristics algorithms [1] given as TDH1, TDH2 and TDH3. Here we are implementing these algorithms for social statistical dataset. We are using relational database which store sensitive information of a set of persons.

TDH1- In order to implement this algorithm, we require some software and hardware requirements. A system, which has computational and processing capabilities with java and with some database compatibility may be used. Here we are using java as front end of processing and MySQL as database.

This heuristic of selecting cuts along minimum bound queries favors queries with smaller bounds. Also, this approach creates imprecision slack in the queries with smaller bounds that could have been used to satisfy bounds of other queries. For running the algorithm TDH1, firstly we have to load the dataset to front end.

Dataset contains following attribute set-

(Age, Education, Work Class, Marital status, Occupation, Race, Gender)

The following table describes these attributes with their specified attribute value.

Table 4.1 various attributes and their values used in working dataset

Attribute Name	Attribute value
Age	Int
Education	Varchar2
Work Class	Varchar2
Marital status	Varchar2
Occupation	Varchar2
Race	Varchar2
Gender	Varchar2

The above dataset is loaded to the front end for generating partitions.

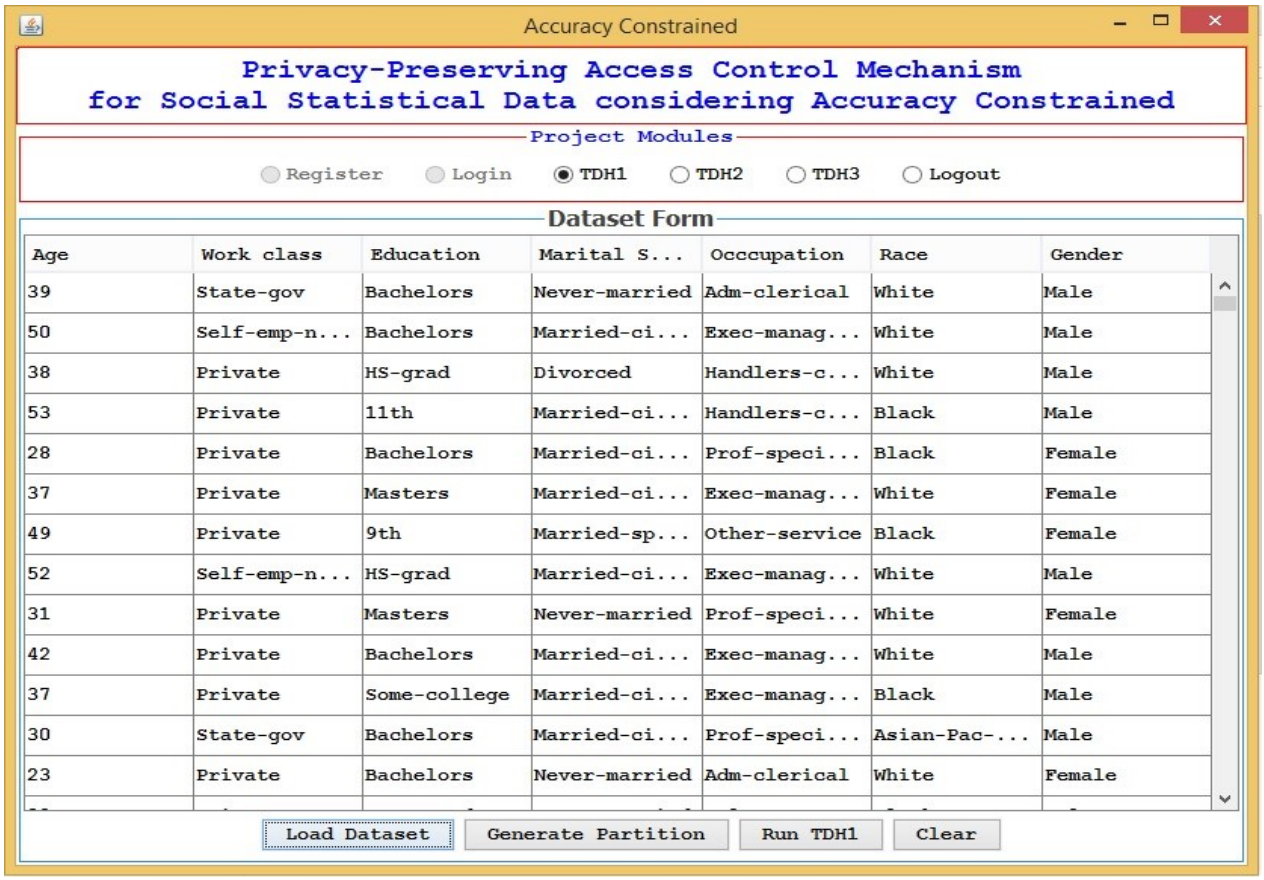
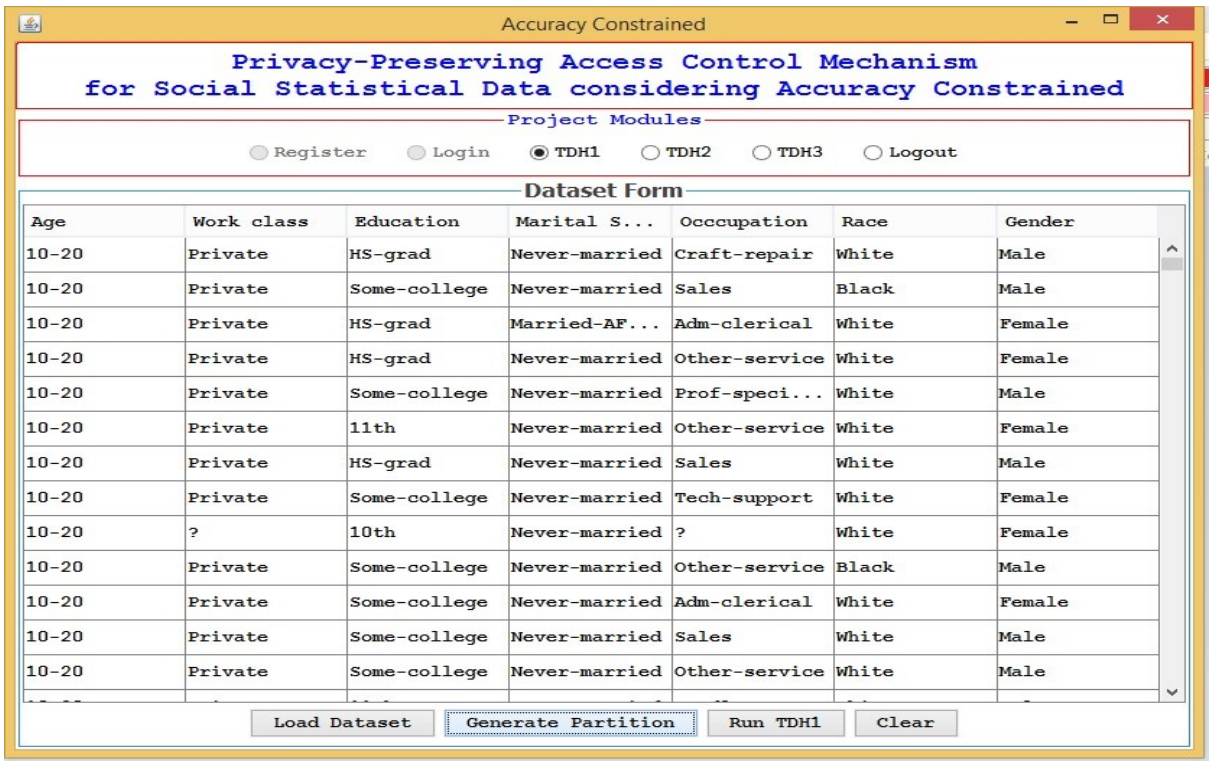


Fig. 4.1 Data set provided to the front end

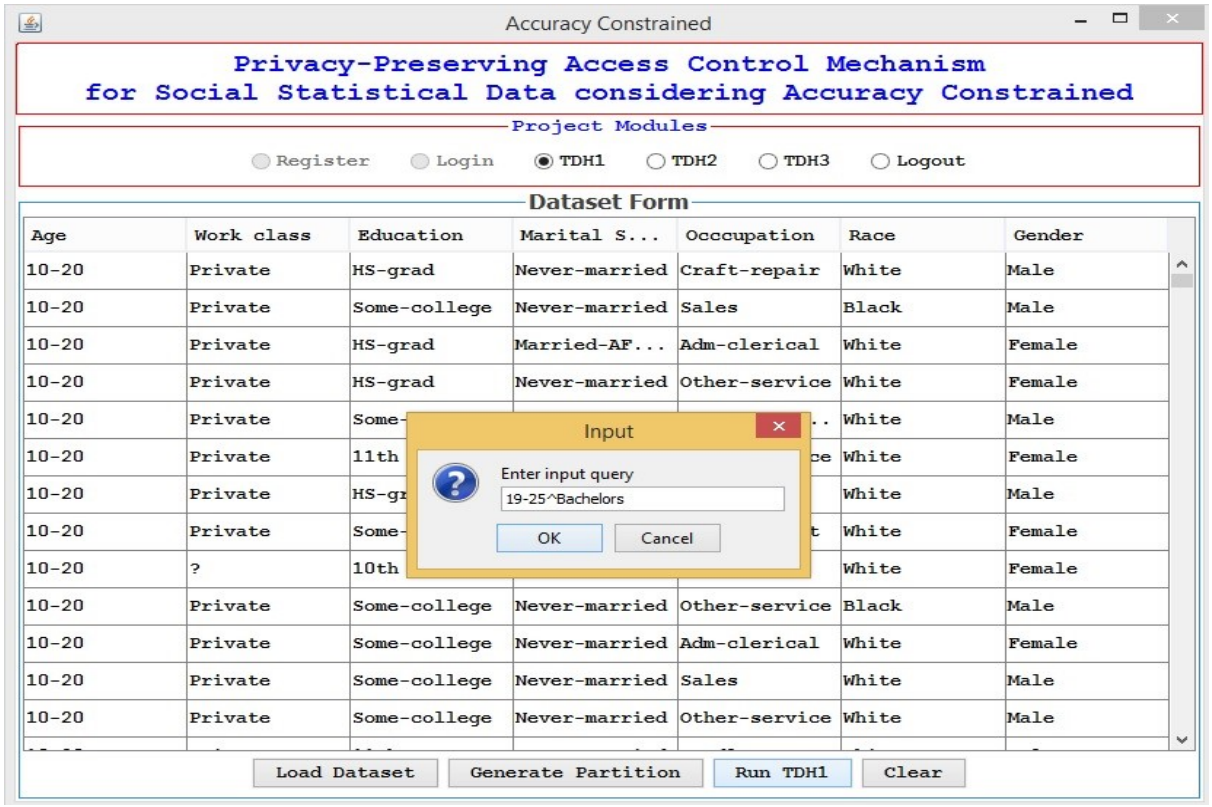
After loading dataset, we generate the partitions for given dataset which actually uses the concept of k-anonymization and sorts the data considering specific range provided.



Age	Work class	Education	Marital S...	Occcupation	Race	Gender
10-20	Private	HS-grad	Never-married	Craft-repair	White	Male
10-20	Private	Some-college	Never-married	Sales	Black	Male
10-20	Private	HS-grad	Married-AF...	Adm-clerical	White	Female
10-20	Private	HS-grad	Never-married	Other-service	White	Female
10-20	Private	Some-college	Never-married	Prof-speci...	White	Male
10-20	Private	11th	Never-married	Other-service	White	Female
10-20	Private	HS-grad	Never-married	Sales	White	Male
10-20	Private	Some-college	Never-married	Tech-support	White	Female
10-20	?	10th	Never-married	?	White	Female
10-20	Private	Some-college	Never-married	Other-service	Black	Male
10-20	Private	Some-college	Never-married	Adm-clerical	White	Female
10-20	Private	Some-college	Never-married	Sales	White	Male
10-20	Private	Some-college	Never-married	Other-service	White	Male

Fig 4.2 Partitioned data set considering Age as QI attribute

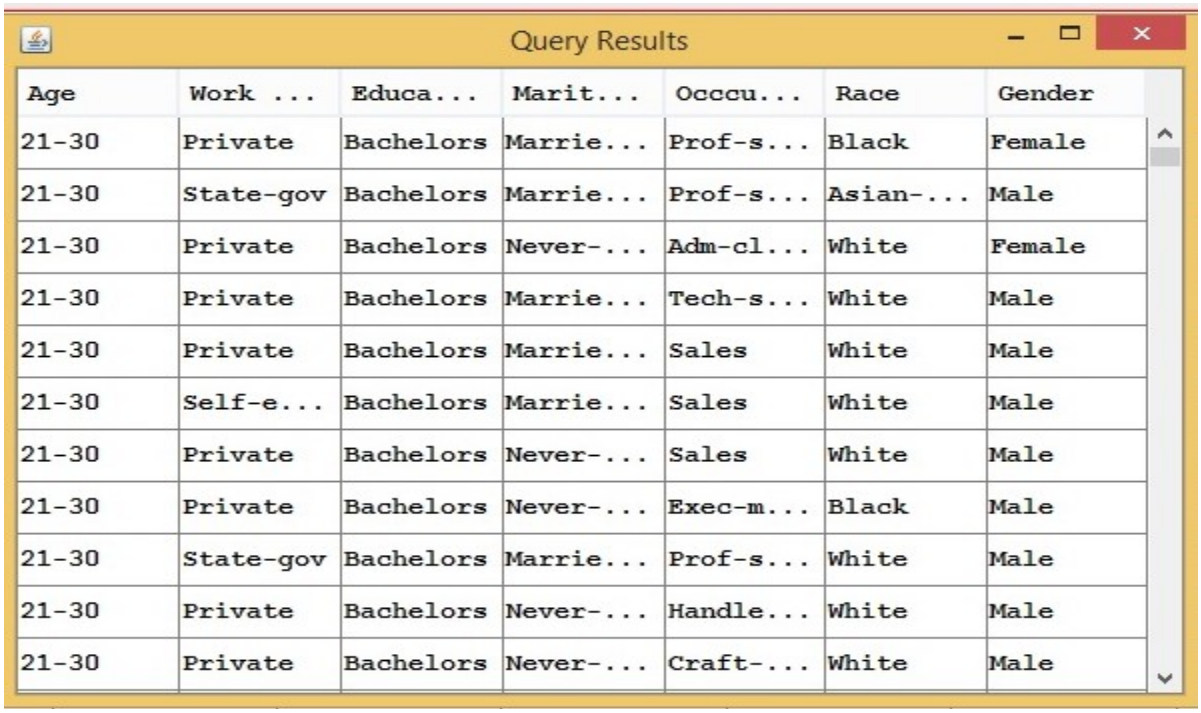
After generating partitions, we can run the algorithm using Run TDH1 button. It will ask the desired query for input which will be according to the specified format as described in access control policy.



Age	Work class	Education	Marital S...	Occcupation	Race	Gender
10-20	Private	HS-grad	Never-married	Craft-repair	White	Male
10-20	Private	Some-college	Never-married	Sales	Black	Male
10-20	Private	HS-grad	Married-AF...	Adm-clerical	White	Female
10-20	Private	HS-grad	Never-married	Other-service	White	Female
10-20	Private	Some-			White	Male
10-20	Private	11th		ce	White	Female
10-20	Private	HS-gr			White	Male
10-20	Private	Some-		t	White	Female
10-20	?	10th			White	Female
10-20	Private	Some-college	Never-married	Other-service	Black	Male
10-20	Private	Some-college	Never-married	Adm-clerical	White	Female
10-20	Private	Some-college	Never-married	Sales	White	Male
10-20	Private	Some-college	Never-married	Other-service	White	Male

Fig 4.3 Query input to partitioned dataset

The result after running the algorithm is shown below.



Query Results

Age	Work ...	Educa...	Marit...	Occcu...	Race	Gender
21-30	Private	Bachelors	Marrie...	Prof-s...	Black	Female
21-30	State-gov	Bachelors	Marrie...	Prof-s...	Asian-...	Male
21-30	Private	Bachelors	Never-...	Adm-cl...	White	Female
21-30	Private	Bachelors	Marrie...	Tech-s...	White	Male
21-30	Private	Bachelors	Marrie...	Sales	White	Male
21-30	Self-e...	Bachelors	Marrie...	Sales	White	Male
21-30	Private	Bachelors	Never-...	Sales	White	Male
21-30	Private	Bachelors	Never-...	Exec-m...	Black	Male
21-30	State-gov	Bachelors	Marrie...	Prof-s...	White	Male
21-30	Private	Bachelors	Never-...	Handle...	White	Male
21-30	Private	Bachelors	Never-...	Craft-...	White	Male

Fig. 4.4 Query result on running TDH1

The result consist the data in anonymized form which is not directly relates to a single person rather than a specific range of persons. It is difficult to identify a person with correct details related to that.

Similarly, we can run **TDH2** and **TDH3** algorithms on considered dataset. There are two differences in TDH2 as compared to TDH1. First one is the traversal of kd-tree which pre-order. Second one is the query bounds which are updated as the partitions are being added to the output (P). The complexity of TDH2 is $O(dIQI^2n^2)$ in time context, which is the same as the time complexity of TDH1. The time complexity of the TDH2 algorithm is $O(dIQI^2n^2)$, which is not scalable for large data sets (greater than 10 million tuples) [2]. TDH3 algorithm is a modified version of TDH2. TDH2 is modified so that the time complexity of O(dIQInlgn) [2] can be achieved, which is done at the cost of increasing imprecision in the query results. For a given partition, TDH3 checks the query cuts only for the query having the lowest imprecision bound. Besides this, it also satisfies the constraint that the query cuts are feasible only when the size ratio of the resulting partitions is not highly skewed. If we are using a skew ratio of 1:99 for TDH3 as a threshold, and if a query cut exists in one partition which size greater than 100 times to the others, then that cut is ignored.

V. CONCLUSION

It is more desirable to have a well designed access control policy to access the database while dealing with sensitive data. It also requires a well designed privacy protection of users in order to ensure the privacy of users. Social media or social networking sites storing private information of users and often more susceptible to privacy breach of users. We are using access control mechanism in which only authorized query predicates are allowed. The privacy preserving mechanisms anonymize the data in order to meet privacy requirements and imprecision constraints on predicates which are defined by access control mechanism (ACM). For data anonymization, we have used k-anonymous Partitioning with Imprecision Bounds (k-PIB). We have used heuristics for partitioning the data to satisfy the imprecision bounds with privacy constraints. In this paper we have used these approaches to work on social statistical data which may concern for some analysis purpose while source of this data may be some social networking sites or some other related databases.

VI. REFERENCES

1. Yung-Wang Lin et. al. “Preserving Privacy in Outsourced Database”, International Journal of Computer and Communication Engineering, Vol. 3, No. 5, September 2014

2. ZahidPervaiz, Walid G. Aref, ArifGhafoor, 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. N. Li, W. Qardaji, and D. Su, "Provably private data anonymization: Or, k-Anonymity Meets Differential Privacy," Arxiv preprint arXiv: 1101. 2604, 2011.
4. "Database access control & privacy: Is there a common ground?" Proc. Fifth Biennial Conf. Innovative Data Systems Research (CIDR), pp. 96-103, 2011.
5. "Ireduct: Differential Privacy with Reduced Relative Errors," by X. Xiao, G. Bender, M. Hay, and J. Gehrke; Proc. ACM SIGMOD Int'l conf. Management of data, 2011
6. "A Framework for Efficient Data Anonymization Under Privacy and Accuracy Constraints," by G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis; ACM Trans. Database Systems, vol. 34, no. 2, article 9, 2009.
7. K. LeFevre, D. DeWitt, and R. Ramakrishna, "Workload-Aware anonymization techniques for large-scale datasets", ACM Trans. Database Systems, vol. 33, no. 3, pp. 1-47, 2008.
8. "Syndromic Surveillance Practice in the United States: Findings from a survey of state, territorial, and selected local health departments" by J. Buehler, A. Sonricker, M. Paladini, P. Soper, and F. Mostashari; Advances in Disease Surveillance, vol. 6, no. 3, pp. 1-20, 2008.
9. 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.
10. 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. .
11. K. LeFevre, D. DeWitt, and R. Ramakrishna, "Mondrian Multidimensional K-Anonymity," Proc. 22nd Int'l Conf. Data Eng., pp. 25- 25, 2006.
12. 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.
13. 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.
14. A. Meyerson and R. Williams, "On the complexity of optimal k-Anonymity", Proc. 23rd ACM SIGMOD-SIGACT-SIGART & principles of database systems, pp. 223-228, 2004.
15. 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.
16. "Proposed NIST Standard for Role-Based Access Control" by D. Ferraiolo, R. Sandhu, S. Gavrila, D. Kuhn, and R. Chandramouli; ACM Trans. Information and system security, vol. 4, no. 3, pp. 224- 274, 2001.
17. P. Samarati, "Protecting Respondents' Identities in Microdata Release", IEEE Trans. Knowledge and data Eng., vol. 13, no. 6, pp. 1010-1027, Nov. 2001.
18. R. Sandhu, Q. Munawer, "The Arbac99 model for administration of roles," Proc. 15th Ann. Computer security applications Conf., pp. 229-238, 1999.