IMPROVING PRIVACY PRESERVING FOR LOCATION DIVERSITY USING K-ANONYMIZATION APPROACH

[1]Thakor Urvashi .K, [2]Prof.Tushar J.Raval
[1]Students Of Masters of Computer Engineering, [2]Associate Professor
[1]Department of Computer Engineering,
[1]L.D College of Engineering, Ahmedabad, India

Abstarct: The rise of mobile technologies are in lead to leverage large amount of personal location information. Personal Information is privacy concern. K-anonymity used for real dataset works on lack of diversity in sensitive regions. We propose making sensitive attribute anonymize with better approach by making low purturbation (minimizing variation because of the process) and ensuring data accuracy and preservation should not suffer so that minimum information loss can be achieved.Privacy-preserving technique for static data release that relies on map anonymizations, rather than trajectory grouping. The proposed model first creates a modified K-anonymity model which identify sensitive attributes then apply k-means clustering on original dataset and and annoymised dataset using approach of perturbation so we can achieve data accuracy and privacy should be preserved . Its aim is to reduce the resolution of spatio temporal information and introduce uncertainty in sensitive information of the user. To propose personal confidentiality, a protection metric for trajectory databases, that protects privacy of the users through diversification of locations

***IndexTerms* - Anonymization, Generalization, Suppression, k-anonymity, Pertubation,K-annonymity**

I.Introduction

1.1 Motivation

The rise of mobile technologies are in lead to leverage large amount of personal location information. Personal Information is privacy concern. K-anonymity used for real dataset works on lack of diversity in sensitive regions. We propose a privacy of confidentiality that ensures location diversity by Limiting probability of user visiting a sensitive location or probabilistic analysis based on adversary knowledge. The map anonymization as a anonymize trajectories and probabilistic methods to improve diversified trajectory and location show to be satisfied diversification effectively So, before share data with other party privacy preserving approach applied on that data and hide some content of that data by which no one can get sensitive information of individual person.

1.2 Problem Statement

One of the issue in Anonymization problem is information loss due to generalization of data it will make data in more general form so due to this information loss problem arise.

1.3 Problem Description

In anonymization approach generalization and suppression process apply on quasi attributes. Due generalization process for making anonymize record in other k-1 records it arises information loss. The proposed model first take the land set dataset and apply modified k-anonymity technique which makes sensitive attribute anonymised with better approach by making low perturbation(minimizing the variation because of process)and ensuring data accuracy such that privacy should be preserved.at last k-means clustering used for better performance.

1.4 Research Objectives

- making sensitive attribute anonymize with better approach by making low perturbation (minimizing variation because of the process)
- ensuring data accuracy and preservation should not suffer so that minimum information loss can be achieved.

1.5 Scope of Dissertation Area

This approach cannot stop Homogeneity attack and background knowledge problem can be solved by l diversity problem and if dataset have multiple sensitive attribute then it will create some new issues. Our anonymization approach works on the map and distorts just the sensitive portions of the trajectories.

II.Background Knowledge

2.1 Data Mining

Data mining also called as knowledge discovery in databases (KDD). It is distinct as the progression of estimating motivating, beneficial and unseen outlines from large bulks of data goods and finds the dealings among the outlines. Data mining job involves utilities fir mathematics data and Artificial Intelligence systems (AI). AI systems includes neural networks and machine learning sometimes one can combine them with database managing system for estimating or evaluating the huge sizes of digital data, which is the resultant form of data sets.

2.1.1 Data Mining Process

Data is poised from several causes in Data range step. Next, Data will be pre-processed by selling with null values and unformatted values. Then, Data will be changed to suitable format which is appropriate for data mining action. Now, Information will be take out from data stock which is unknown but data mining. Finally, estimation of outlines for choice construction takes place.

The explicit area of data mining method is to pull out the unseen data from a data set and revolution it into a good comprehensible structure for upcoming use.

Phases of Data mining process consist of following

1. Database and data base association state
2. Data pre-processing
3. Inference conclusions
4. Identifying relationships among the patterns
5. Complexity in analysing the pattern
6. post-processing of discovered patterns for report generation
7. Online updating

2.2. Privacy Preserving Data Mining

Privacy conserving data mining goals to offer valid data mining consequences by not skimpy the underlined sensitive data. Data mining methods extracts valued information from data stores. When the techniques are applied, it not only extracts useful data, may also reveal sensitive information. So as to offer defense for sensitive data some privacy conserving methods can be applied on unique data then mining can be achieved.

1. Information mining reasons community and principled problem, because it reveals information which would requires secrecy?
2. Privacy preserving data mining provides safety to private information in contradiction of unofficial access is a long-standing achievement aimed at data mining safety study civic and for the governmental works.
3. Later, the safekeeping subject is one of the emerging area that became valuable research area in data mining.

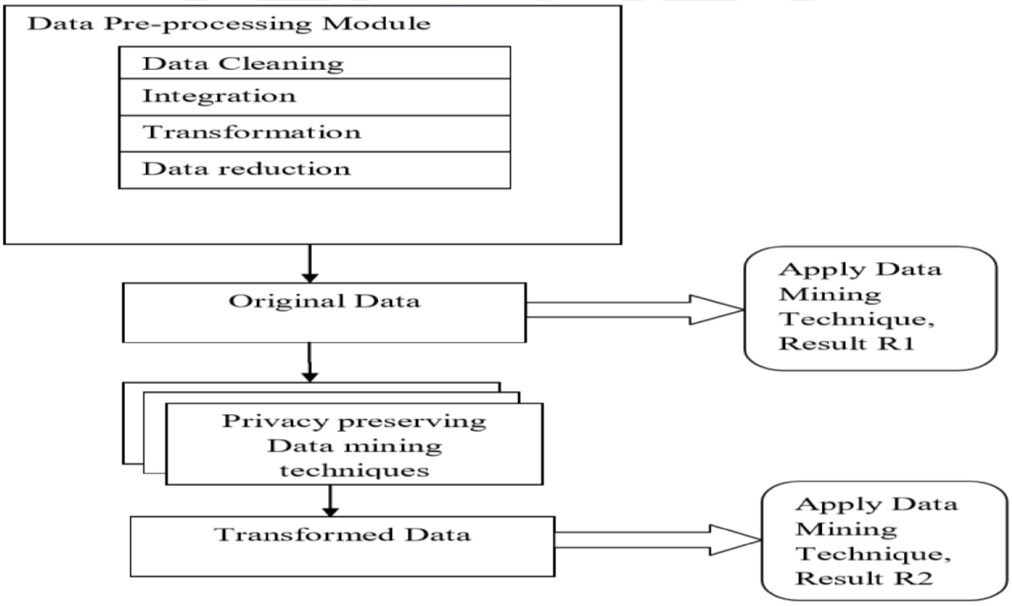
Data Pre-processing Module
Data Cleaning
Integration
Transformation
Data reduction
Original Data
Apply Data Mining Technique, Result R1
Privacy preserving Data mining techniques
Transformed Data
Apply Data Mining Technique, Result R2

Fig. 2.2 Privacy Preserving Data Mining

2.3 K Anonymity

In several applications, the data histories can be finished ready-to-access by just eliminating important recognizers, like person name and Aadhaar or Voter ID, or social security numbers from personal records. But, even then, we can use other types of features (known as pseudo-identifiers) to properly recognize the records. For instance, attributes such as age, pin code and gender are offered in a variation of free tuples such as the censuses. If these attributes are similarly available in a given data set, they may be used to identify the identity of the person the record corresponds to. One of the key Observations is when the more attributes are available to public, it is easy to identify the identities with the group of attributes [4][5]. In The k-anonymity methods offers privacy and introduces an unidentified record via generalize and/or suppress. data in situation of generalization, the values in that database are swapped with some connected values. For sample, if data values for the Age attribute in the dataset are 31, 32, 33, 34, 35 and 36, then they can be represented as (31-36). On the further pointer, in the situation of suppression, the values in a database are concealed or deleted. For instance, the suppressed value might be denoted as 3* for the real values 31, 32, 33, 34, 35 and 36 in records. However, generalization is well as related to suppression [6], since the generalization discloses at lesser some information as related to suppression. But, the anonymous dataset made by generalization and/or suppression consequences in info loss. design of info loss is as trails Study a medical group that needs to offer the medical database to the study work for the medical study purpose. The medical dataset (Table 1) contains of 3 types of attribute

1). Identifier (ID),

2). quasi-identifier (QI)

3). sensitive attribute (SA).

Usually, the identifier attribute takes uninvolved from the record to preserve privacy of specific. But, it is not sufficient to eliminate the only identifier attribute to defend the privacy of separate. The other attributes such as Zip code, Age, Gender are considered as quasi-identifier (QI). quasi-identifier attributes classified in 2 types. Numeric and Categorical. The attribute Disease is measured as a sensitive attribute (SA). In this Table 1, the attribute Age and Zip code are considered as numeric attribute where, the attribute Gender is measured as categorical attribute.

III.Literature Review

Privacy Preserving for Spatio-Temporal Data Publishing Ensuring Location Diversity Using K-Anonymity Technique[1]
privacy-preserving technique for static data release that relies on map anonymizations, rather than trajectory grouping
The rise of mobile technologies are in lead to leverage large amount of personal location information. From knowledge discovery in different point of view, these data are usable, but that personal information is the privacy concerns. There exist many algorithms in the literature namely, perturbing, suppression, generalizing their data to satisfy privacy, required by individuals. Current techniques try to ensure distinguishability between trajectories in real dataset. K-anonymity used for real dataset works on lack of diversity in sensitive regions. We propose a privacy of confidentiality that ensures location diversity by limiting probability of user visiting a sensitive location or probabilistic analysis based on adversary knowledge.Anonymizing trajectory with underlying map, that is interest of point create confusion areas around sensitive locations. Then use map anonymization as a anonymize trajectories and probabilistic methods to improve diversified trajectory and location show to be satisfied diversification effectively

K-Anonymity: A model for protecting Privacy[2]
Re –identification of data using K-anonymity Protection model.
Society is experiencing exponential growth in the number and variety of data collections containing person-specific information as computer technology, network connectivity and disk storage space become increasingly affordable. Data holders, operating autonomously and with limited knowledge, are left with the difficulty of releasing information that does not compromise privacy,confidentiality or national interests. In many cases the survival of the database itself depends on the data holder's ability to produce anonymous data because not releasing such information at all may diminish the need for the data, while on the other hand, failing to provide proper protection within a release may create circumstances that harm the public or others.

Anonymous Usage of Location-Based Services through Spatial and Temporal Cloaking[3]
This approach location-based services collect and use only de-personalized data—that is, practically anonymous data.The more difficult issue is decoupling the anonymizer from the current client-server architecture. For individual users to remain anonymous, the location server must have sufficient users within a geographic locale; unless the different users subscribe to the same location service.Advances in sensing and tracking technology enable location based applications but they also create significant privacy risks. Anonymity can provide a high degree of privacy, save service users from dealing with service providers' privacy policies, and reduce the service providers' requirements for safeguarding private information. However, guaranteeing anonymous usage of location-based services requires that the precise location information transmitted by a user cannot be easily used to re-identify the subject. This paper presents a middleware architecture and algorithms that can be used by a centralized location broker service. The adaptive algorithms adjust the resolution of location information along spatial or temporal dimensions to meet specified anonymity constraints based on the entities who may be using location services within a given area. Using a model based on automotive traffic counts and cartographic material, we estimate the realistically expected spatial resolution for different anonymity constraints. The median resolution generated by our algorithms is 125 meters. Thus, anonymous location-based requests for urban areas would have the same accuracy currently needed for E-911 services; this would provide sufficient resolution for wayfinding, automated bus routing services

Protecting Privacy in Continuous Location-Tracking Applications[4]
This paper introduced disclosure control algorithm that hide user Position in sensitive area. Location based tracking

mechanism is implemented. Recent technological advances in wireless location tracking (such as that found in cell phones1Recent technological advances in wireless location tracking (such as that found in cell phones1and radio frequency identification (RFID) chips, among others) present unprecedented opportunities for monitoring individuals' movements. While such technology can support useful location-based services (LBS), which tailor their functionality to a user's current location, privacy concerns might seriously hamper user acceptance. LBSs can be classified into three types: position awareness, sporadic queries, and location tracking. Position awareness refers to devices that monitor an individual's position, such as in-car navigation systems or GPS-enabled PDAs, but that only use such information internally. Sporadic queries apply to services in which an individual initiates the transfer of position information to an outside service provider. These queries contain only the user's current position—as in point-of-interest queries to find the nearest hotel, for example. Finally, location-tracking services

receive frequent updates of an individual's position— for example, experimental automotive telematics applications, which seek to improve transportation through the use of information technology, use such updates to estimate highway gridlock and reroute drivers around traffic jams.Continuous location-tracking applications exacerbate privacy problems because individuals' desired levels of privacy can be situation-dependent.

Anonymization Of Data Using Mapreduce On Cloud[5]
In this work highly scalable "Two Phase Top Down Specialization approach" is used for anonymization of data based on MapReduce on cloud. Specialization is required in an anonymization process to make effective use of parallel computing capability of MapReduce on cloud. This process is split into two phases. In 1st phase the original data set is partitioned into different groups of smaller data sets and these smaller data sets are anonymized in parallel by the help of MapReduce to producing intermediate results. In 2nd phase the intermediate results which are generated in first phase are integrated into one data set, and this data set is further anonymized to achieve consistent k-anonymous data set. This approach tends MapReduce to achieve the concrete computation in both the phases. To do specializations on cloud data sets, a group of MapReduce jobs are carefully constructed and coordinated.

IV Proposed Approach for Anonymization

4.1 Overview of Methodology

1. Input database take the input statistics.
2. Count Stastics in database.
3. Apply cleaning Process on database.
4. Partition the column according attributes.
5. Store Sensitive Data to a Separate table
6. Covert Quasi Identifiers (Qi).
7. Call Modified K-anonymity method
8. Apply K-Means Algorithm on Original Dataset.
9. Apply K-means Algorithm on Anonymised Dataset.
10. Display the result.

4.2 Algorithm: Modified K-anonymity for Location Based Spatial data

Input: a releasable dataset D and a diversity threshold value l.

Output: a releasable dataset D', which ensures that each EC has the same sensitive attribute values set before and after update and has minimum information loss.

Steps:

```
Step 1: Input db.
Step 2: Count Statistics
Step 3: cleaning data process
Step 4: partition columns (sensitive and no sensitive attributes)
Step 5: store sensitive data to a separate table (possibly hash values)
Step 6: Convert the Quasi-identifiers into semantic Hierarchical tree for classification.
Step 7: perform modified k-anonymity
Step 8: Form initial equivalence class, S= {E1, E2, E3… En}
While T exists an equivalence class of size <k do
Select an equivalence class Ei of size <k;
Calculate the pair wise distance between Ei and remaining equivalence classes in T,
Find the equivalence class Ej with the smallest distance to ECi.
Do till classes found
```

Step 9: Apply k-Mean clustering algorithm with different values of k on original dataset D having sensitive attribute a.
Step 10: Apply k-Mean clustering algorithm with different values of k on Anonymised dataset D' having sensitive attribute a.display results.

K-means Clustering:
k-means clustering algorithm is a data mining and machine learning tool used to cluster observations into groups of related observations without any prior knowledge of those relationships. By sampling, the algorithm attempts to show in which category, or cluster, the data belong to, with the number of clusters being defined by the value

4.2.1 Proposed System Flow

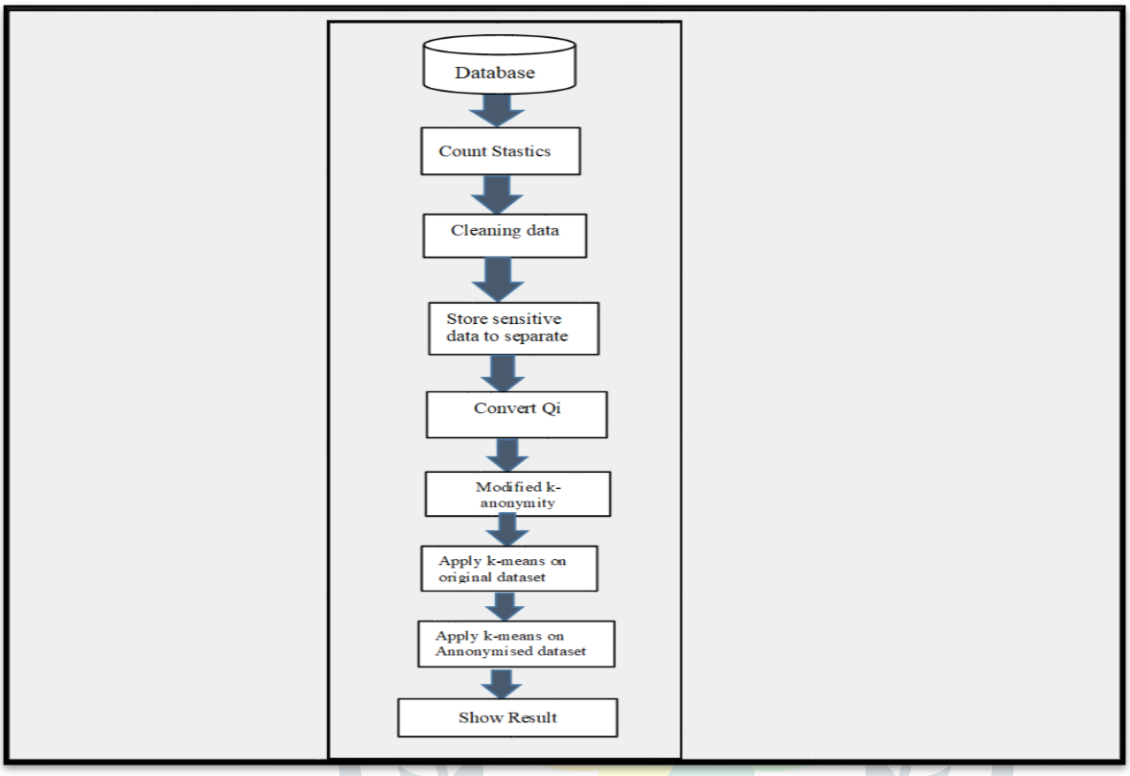
Database
Count Stastics
Cleaning data
Store sensitive data to separate
Convert Qi
Modified k-anonymity
Apply k-means on original dataset
Apply k-means on Annonymised dataset
Show Result

Fig 4.2.1 Proposed System Flow diagram.

This diagram shows the proposed system architecture and system process flow.in first step we are taking the dataset as input and counting the total number of records in dataset which process is called is called count statistics then after applying data mining cleaning process on data to remove duplicated dirty data and in next step we are taking separate table to store sensitive data and converting Qi quasi identifiers .after converting quasi identifier we are applying modified k-anonymity method this method decide the class according to range and create the pair of class. Then apply k-means clustering on original dataset and k-means clustering method on annonimysed dataset for creating different cluster and finally we are displaying final result.

V tools and technology

1. Net Beans IDE 7.3.1: This tool used for actual programming in java language. Net Beans IDE as the original free Java IDE. It is that, and much more! The Net Beans IDE provides support for several languages (PHP, JavaFX, C/C++, JavaScript, etc.) and frameworks.
2. JDK 1.8: The Java Development Kit (**JDK**) is a software development environment used for developing Java applications and applets. It includes the Java Runtime Environment (JRE), an interpreter/loader (java), a compiler (javac), an archiver (jar), a documentation generator (javadoc) and other tools needed in Java development.
3. Servlet
4. Weka tools
5. Jsp(java server page)

VI Datasets Used

1. **Landsat:**

The Landsat satellite (stat log) data is one of the many sources of information available for a scene. The interpretation of a scene by integrating spatial data of diverse types and resolutions including multispectral and radar data, maps indicating topography, land use etc. is expected to assume significant importance with the onset of an era characterized by integrative approaches to remote sensing (for example, NASA's Earth Observing System commencing this decade). So we are comparison our results with 6495 trajectories in datasets. This dataset is available over UCI machine learning repository which is having variation of the multi-spectral values of pixels in 3x3 neighbor hoods in a satellite image, and the classification associated with the central pixel in each neighborhood. The aim is to predict this classification, given the multi-spectral values. Here, preservation of privacy in this database is required because it gives information about the land which will reveal the geographic information. It is the data or information that identifies the geographic location of features and. It requires for regular monitoring and updating of security plan boundaries on Earth, such as natural or constructed features, oceans, and more. Spatial data is usually stored as coordinates and topology, and is data that can be mapped

VII Equation for Precision and Recall

True class A (TA) - correctly classified into class A
False class A (FA) - incorrectly classified into class A
True class B (TB) - correctly classified into class B
False class B (FB) - incorrectly classified into class B

Precision = TA / (TA + FA)
recall = TA / (TA + FB)

VIII Experimental Results and Discussion

The performance of our approach is measured in Precision and Recall .All experiments were done on 3.20-GHz intel core i5 CPU running on Microsoft Windows 8. Constituent data is stored as a relational database using MySQL.
In experiment to use real map as using latitude and longitude in Pune schools dataset as excel file that assign direction of vertex edges information in graph. Implementation that works as begin directions in obtain path from any node in graph to any another node. If they are picks all nodes on the path a creates trajectory dataset. To measure output of systems using Precision and Recall parameter. In Table 8.3 and 8.4 present experimental values of precision and recall to find output graph of system in fig 2

IX Comparison Graph

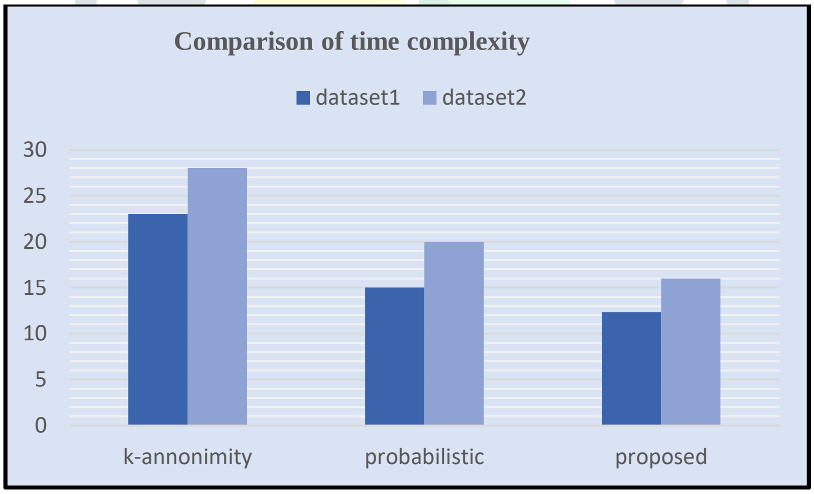

Fig 8.1 Comparison of time complexity

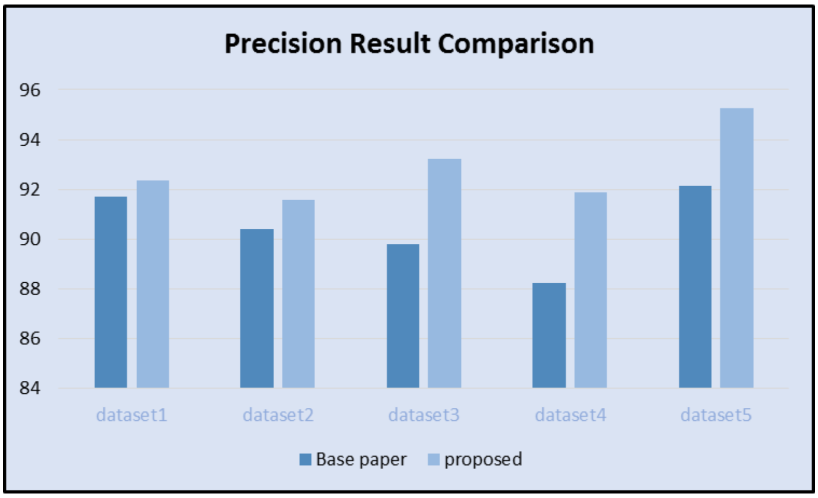

Fig 8.2 Comparison of Pricision

X.Conclusion

I have addressed the issue of information loss which is solved at a good extent with proposed approach .so far I have acquired result through my proposed methodology. These results will be analyzed with respect to the results of my base papers implementation

XI References

[1] L. Sweeney, "k-Anonymity a model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems.

[2]"Ensuring location diversity in privacy-preserving spatio-temporal data publishing" Author:- A. Ercument Cicek · Mehmet Ercan Nergiz · Yucel Saygin

[3] "Anonymous Usage of Location-Based Services through Spatial and Temporal Cloaking "

[4] "Protecting Privacy in continues Location Tracking Application".- MarcoGruteser, Xuan Liu

[5] "Anoymization of data using map reduce on cloud."Author:-Mallappa Gaurav, N. V. Karekar, Manjunath Suryavanshi

[6] Pawan R. Bhaladhare And Devesh C. Jinwala, "Novel approaches for privacy preserving data mining in k-anonymity model"

[7] Ninghui li tiancheng li "T-closeness privacy beyond k-anonymity and l-diversity" department of computer science, purdue university

[8] "Enhancing privacy preservation in data mining using cluster based greedy method in hierarchical approach" Indian journal of science and technology, 2016.

[9] J. Lin and m. Wei, "an efficient clustering method for k-anonymization," in proceeding of international workshop on privacy and anonymity in information society, 2008.

[10]R. C. Wong, j. Li, a.w. Fu, and k. Wang, "(l, k) anonymity an enhanced k-anonymity model for privacy preserving data publishing," in proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining, 2006.

[12] Denning d. Secure statistical databases with random sample queries. Acm tods journal, 5(3), 1980.

[13]Jagannathan g., Wright r. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. Acm kdd conference, 2005.

[14]Jagannathan g., pillaipakkamnatt k., wright r. A new privacy- preserving distributed k-clustering algorithm. Siam conference on data mining, 2006.

[15]Aggarwal g., feder t. Kenthapadik., motwani r., panigrahy r. Thomas d. Zhua. Approximation algorithms for k-anonymity. Journal of privacy technology, paper 20051120001, 2005.

[16]Gedik b., liu l. "A customizable k-anonymity model for protecting location privacy", icdcs conference, 2005.

[17]M. E. Kabir, h. Wang and e. Bertino, "efficient systematic clustering method for k-anonymization," act a informatics, 2011.