International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 10, Issue 4 (April 2014), PP.50-60

Enhanced Anonymization and Clustering of OSN

Utkarsh, K. Senthil Kumar
Department of Computer Application, SRM University, Chennai.
Assistant Professor (S.G). Department of Computer Application, SRM University, Chennai.

Abstract:- In current scenario online social networks such as Facebook, LinkedIn are increasingly utilized by many people. These platforms have successfully created a virtual world by allowing the user to connect with each other just by sharing information about themselves. The information disclosed in here are public as well as private which has created a lot of trouble in recent times. The More data is shared by people the more privacy issues are being violated. In shared data publishing process, we need not only protect the privacy of data but also insure the data's integration. To address this issue we proposed the solution of reliable and secure data set by using various Anonymization and clustering concepts and then this paper explores data randomization, i.e. our method maintains statistical relations among data to preserve knowledge, whereas in most anonymization methods, knowledge is lost then we shift on how to launch inference attacks using released social networking data to predict private information and then our work shows the effectiveness of these techniques and finally we worked on how we can decrease the privacy leakage from social networks**.**

Keywords:- K-Anonymity, Quasi-Identifiers, I-Diversity, Generalization, Attacks and Privacy Model, NBDH**.**

I. INTRODUCTION

In recent year online social networking applications have gained a dramatic attention for research as due to its wide proliferation in demand. Nearly 58 percent of our population uses social networking sites, referring to the numbers Facebook alone has more than 1.23 billion users[1], Twitter comprises more than 200 million users[2], LinkedIn comprises more than 277 million[3], VK (originally VKontakte) comprises more than 172 million users[4]. As a part of their offering these networks allow people to list details about themselves that are relevant to the nature of the network. An online social network can be defined as the web of network as shown in Figure 1,with enormous numbers of connected Links and nodes which provides a platform to build social relation among people who, for example, share interests, activities, backgrounds or real-life connections. It consists of user profile, his social links, and a variety of additional services.

In the year 2005, a study was performed to analyze data of 540 Facebook profiles of students enrolled at Carnegie Mellon University[5]. It was revealed that 89% of the users gave genuine names, and 61% gave a photograph of themselves for easier identification. Majority of users also had not altered their privacy setting, allowed a large number of unknown users to have access to their personal information (the default setting originally allowed friends, friends of friends, and non friends of the same network to have full view of a user‗s profile). It is possible for users to block other users from locating them on Facebook, but this must be done by individual basis, and would therefore appear not to be commonly used for a wide number of people. Most users do not realize that while they make use of the security features on Facebook the default setting is restored after each update. All of this has led to many concerns that users are displaying far too much information on social networking sites which may have serious implications on their privacy. Here a question arises that what exactly do we mean by privacy in social networking? Privacy an isolated state in which one is neither observed nor disturbed by other people i.e.; the user's information must not be misused and they should not be disturbed by the other people.

Twitter has admitted that they have scanned and imported their user's phone contacts onto the website database in order to learn more about their users[6]. Most users were unaware that Twitter has created this way for new users to search for their friends. Twitter has stated that they will have their privacy guidelines illustrated more clearly in the future. More than 1,000 companies are waiting in line to get access to millions of tweets from users that are using the popular social networking website[7]. The more data is shared by the people the more privacy issues are being violated.

In a social network privacy concern of individuals are of two types privacy breach after data is released and other is leakage of private information. The best example of privacy breach is the AOL search data scandal[8]. In the year 2006, AOL released a compressed text file on one of its websites containing twenty million search keywords for over 650,000 users over a 3-month period, intended for research purposes however; the public release meant that the entire Internet could see the results. AOL themselves did not identify users in the report; however, personally identifiable information was present in many of the queries and as the queries were attributed by AOL to particular user accounts, identified numerically, an individual could be identified and matched to their account and search history by such information. Before we go on for further discussion we must know what is personally identifiable information (PII)? PII is specific category of particularly sensitive data that includes an individual's unencrypted first name or initial and last name, in combination with any one or more of the following like Social Security number, Drivers license number, Financial account number, credit card number, Debit card number and etc.

Companies believe that by using data mining technologies they would be able to gather important information that can be used for marketing and advertising. In data mining, there are two conflicting goals: privacy protection and knowledge preservation privacy. On the one hand, we anonymize data to protect privacy; on the other hand, we allow miners to discover useful knowledge from anonymized data.

This paper focuses on the problem of private information leakage from social network of an individual and then we have proposed a new technique for solving this type of problem. Anonymization and clustering concepts have been used and an experimental model has been setup that elaborates how we will solve the problem of privacy? How to secure the information that is being leaked. In our model we have used enhanced k-Anonymity and clustering concept which store the data in a separate database table for secure and reliable reference. Clustering will seek the actual tradeoffs between data utility and privacy leakage.

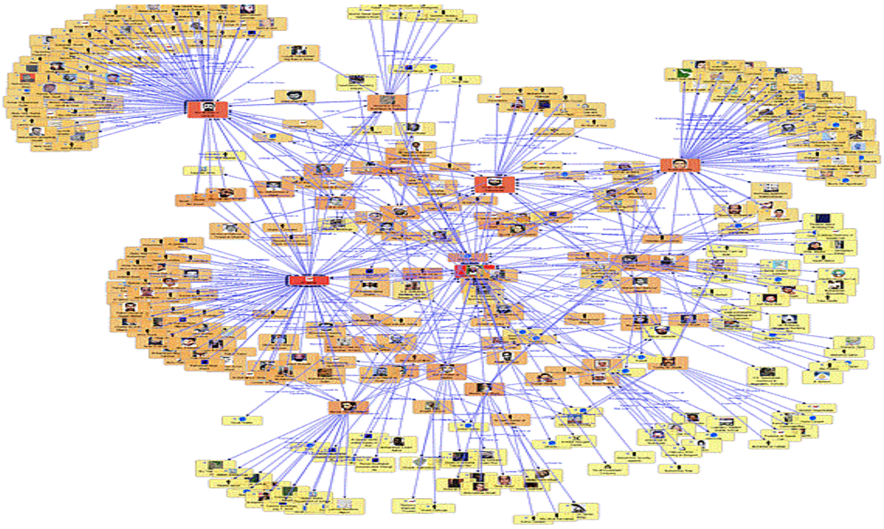


Figure 1 Links and nodes in Social Networking

II. RELATED WORK

Security aspects in social networks are getting more and more attention in the recent years and hence this paper explores all the new possibility of securing the data and information that are shared by the users in social network. The limited capacities of social networking security nodes and the complex algorithms of the security protocols make the subject challenging. In order to provide security in social networks, communications should be encrypted and authenticated. The main issue is how to set up secret keys between nodes to be used for the cryptographic operations which are known as the key agreement.

Unfortunately, security is in general considered to be expensive. Its cost is even more noticeable in social network due to the limited resources. Thus, in order to provide a sufficient level of security while properly utilizing the available resources, it is important to have a good understanding of both the cost and the features of the security algorithms used. For example, if a sensor device does not have enough available memory to run a specific security protocol, it might be better to use an alternative algorithm which requires less memory but might be more secure. This work's contributions are threefold.

First, we have analysed all the anonymization technique and then we have studied how the different algorithm parameters (e.g. key size) influence the privacy concern. Further, based on our study, we have evaluated the tradeoffs between clustered security algorithms. The second contribution is the proposing of a new algorithm which we have written as an experimental setup for Facebook users in java where we have shown how to anonymize the data set and how to store the data with the concept of clustering. Chang et al. [9] and Law et al [10] have shown several ways of anonymizing social networks. However, our work focuses on concluding from evidence and reasoning rather than from explicit statements details from nodes in the network. In Backstrom et al. [11] their work mainly focuses on the prediction of the private attributes of users in four different domains which are Facebook, Flicker, Dogster, and BibSonomy. They do not attempt to anonymize or clean any graph data. They have also measured the memory usage of the standard modes of operation but have not evaluated their energy consumption. In addition, they do not specify the platform they use in their experiments. Compared to their work, we provide a more detailed evaluation of both security algorithms and modes of operation.

As we mentioned earlier, none of the previous work has studied the impact of the different algorithm Further, no previous work has evaluated the. Other papers have tried to infer private information inside social networks. Het et al [12] focused on various ways to deduce private information via friendship links by creating a network from the links inside a social network while they have used theoretical attributes to analyze their learning algorithm.

The existing work does scrutinize the model very briefly and analyzes access control for all the shared data in online social networking sites. The problem of leakage of private information from online social networking is still a critical issue. In Proposed System we have implemented a new concept in Facebook i.e., proof-of-concept for the collaborative management of shared data, we have also tried to show how the data are shared and how we can set the privacy. Our prototype application enables multiple associated users to specify their authorization policies and privacy preferences to co-control a shared data item. We have show how the online social network data could be used to predict some individual private detail that a user is not willing to disclose and then we have proposed a new mechanism of clustering that can be used for classifying and storing of the databases.

III. PROPOSED TECHNIQUES

Data anonymization is the process of destroying the tracks of the data so that the origin of the data cannot be retained back. It converts the text into non readable non human form by applying various encryption and decryption technique and for this we have enlightened our work on Anonymization technique and have used enhanced K-Anonymity as our basic tool, enhanced in a way that we have expanded the definition and scope of k-Anonymity in our experimental setup and have used it with combination of clustering. K-Anonymity provides syntactic guarantees for data i.e.; they make sure that an individual cannot be identified from the data but does not consider inference attacks that can be launched to infer private information. Basically, it guarantees that the change in one record does not change the result too much. So clearly this does not help us to make an accurate data mining model that can predict sensitive information. In current scenario many different algorithms have been developed that has similar working with k-Anonymity but they all lacks in certain parameter.

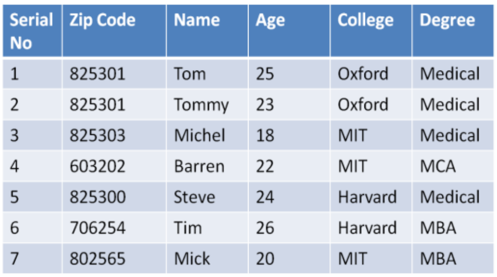
Our main motive is to redefine social networking sites with preventing of data set through data mining techniques. To begin our work we first need to understand two basic things which are used for the formal definition of privacy. First, we clearly need to understand how to protect out data set when the hackers already knows where are all the hidden and unhidden private information related to users. Second, we need to analyze very thoroughly that if inference attack is done based on user’s background information then how we will protect our data when hackers succeeded to enter into our system. For example, if the user has disclosed zip code where he lives and based on it, can an outsider predict his/her political interest? Well to solve this problem we came up with a new definition of k-Anonymity where we tried to manipulate all the related links by three ways: adding details to nodes, modifying existing details and removing details from nodes although it is successful against all possible background information but, this goal is not realistic when we have to deal with a large set of data. For example if opponent has a background information stating that tom’s education is the same as the majority of people in California have then any aggregate statistics can be used for an inference attack. In order to solve this issue we switched from absolute database to relative database.

During our work we found that using relative database was much more reliable as compared to absolute database. To address the second issue listed above; we need to estimate the performance of the best classifier that can be built by using the released social network data and the adversary’s background knowledge. Therefore, in our privacy definition, we try to explore the additional scope of k-anonymity along with clustering.

The success of algorithm and our experimental setup has been shown using graphical representation. In this paper we have tried to develop a relative new privacy definition based on the difference in classification accuracy with and without the released social network data for a given background definition. We would like to state clearly that our privacy definition focuses on preventing inference attacks only and could be used with other definitions that tries to protect against other privacy attacks.

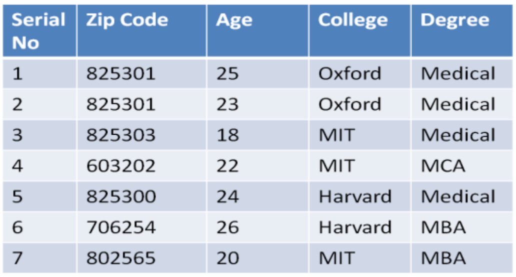
IV. ATTACKS AND PRIVACY MODELS

Generally when people talk about privacy they think to keep the information from not being available to others but on a serious note privacy concern rises when the information is misused by others for negative impact on someone’s life. The problem is that once the information has been released it is impossible to prevent misuse. For example suppose we have raw database as shown in Table 1 which contain detail information about users and similarly another database is available to adversary as shown in Table 2 where the data are incomplete but by linking the adversary can easily extract the information which he wants to know.



Serial No	Zip Code	Name	Age	College	Degree
1	825301	Tom	25	Oxford	Medical
2	825301	Tommy	23	Oxford	Medical
3	825303	Michel	18	MIT	Medical
4	603202	Barren	22	MIT	MCA
5	825300	Steve	24	Harvard	Medical
6	706254	Tim	26	Harvard	MBA
7	802565	Mick	20	MIT	MBA

Table 1 Raw database



Serial No	Zip Code	Age	College	Degree
1	825301	25	Oxford	Medical
2	825301	23	Oxford	Medical
3	825303	18	MIT	Medical
4	603202	22	MIT	MCA
5	825300	24	Harvard	Medical
6	706254	26	Harvard	MBA
7	802565	20	MIT	MBA

Table 2 Adversary database

As we can see that although if the adversary has this incomplete database he can easily extract his information like using pin code one can easily say that the area with a common pin number of 8253 have maximum interest in medical field and hence they can use this information in many different way to manipulate people. In current scenario the two main privacy preserving paradigm has been established: K-anonymity[13] which prevents identification of records in data and I-Diversity [14] which prevents the association of an individual.

The database is said to be K-anonymous where attributes are generalised until each row is identical with other row in the form of K-1 rows, it thus prevents the database linkage. It ensures and guarantees that the data released is accurate. It basically focuses on two paradigms: generalization and suppression. To protect user's integrity it removes Quasi Identifier. Quasi Identifier is the piece of information or set of attribute that are not unique identifier but when these information are related with other attributes they retrieve themselves to form PII through which 87% of the population of US can be easily identified. In spite of these benefits it has certain limitation such as:

- It lacks hiding property i.e., if an individual is present in database it cannot hide it.
- It doesn't protect the adversary attack based on background knowledge of users.
- It the adversary has already knowledge about k-anonymity he can easily breach the security.
- There is a plenty of data loss evolved in this method. It cannot handle large database.

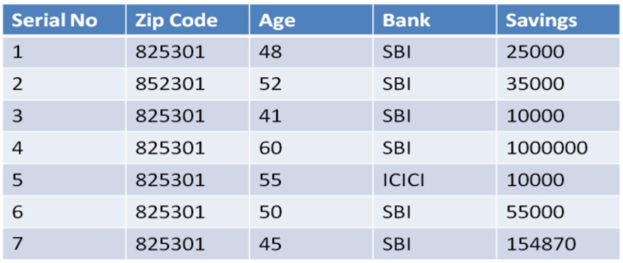
Although it has a very dynamic approach for data protection namely perturbation but the main drawback with this approach is that it cannot tell clearly how much privacy is guaranteed. It lacks a formal definition of privacy. Moving on to the second established concept named I-Diversity which prevents the association of an individual from database but, it also lack in certain parameter.

3.2 I-DIVERSITY

This approach lacks in a particular paradigm in a sense of generalization, suppose we have a group of different record and all those record have unique identifier it is for sure that the attacker will not be able to extract the exact information but the problem which lies here is what if the value in which each individual are interested in are same for every group as shown in Table 3, the most frequently used bank in Delhi is SBI. Now here it would be easy for the attacker to predict the information from this data. Thus this is the basic drawback of I-Diversity although it is the most secure approach we have till now. I-Diversity basically uses two major techniques: Generalization and Permutation. The technique of l-diversity was proposed not only to maintain the minimum group size of k, but also to focuses on maintaining the diversity of the sensitive attributes. The l-diversity model for privacy is defined as follows:

3.3 DEFINITION

Let a q∗-block be a set of tuples such that its non-sensitive values generalize to q∗. A q∗-block is l-diverse if it contains l "well represented" values for the sensitive attribute S. A table is l-diverse, if every q∗-block in it is l-diverse.



Serial No	Zip Code	Age	Bank	Savings
1	825301	48	SBI	25000
2	852301	52	SBI	35000
3	825301	41	SBI	10000
4	825301	60	SBI	1000000
5	825301	55	ICICI	10000
6	825301	50	SBI	55000
7	825301	45	SBI	154870

Table 3 Adversary database

V. PROBLEM SKETECH

Here in this section we have tried to provide a broad overview of the different techniques for preserving privacy in data mining. we have provided a brief review on the major algorithms and approaches available for existing system. Additionally we proposed a modification on K-anonymity and on clustering as well. While working a new combination of algorithm has been proposed where we tried to elaborate why there is need of enhanced K-anonymity with clustering. Working on the major areas we encountered with different types of threat and it was found that there are three main information disclosure threat which are:

- Members disclosure Protection.
- Identity disclosure.
- Attribute disclosure.

To battle privacy attacks and develop protection techniques in social network all anonymization technique have some limitation and for this we have added it with clustering based approach and graph modification mode have also been used.

4.1 ANONYMIZATION TECHNIQUES

Anonymization is widely used technique for securing data it, converts clear text data into a nonhuman readable and irreversible form. Data anonymization ensures a secure transfer of information across a boundary, such as between within any department or between two department. Generalization, perturbation, bucketization are some of the popular anonymization approaches for relational data in social networking.

i) GENERALIZATION

Generalization is one of the commonly anonymized approach, which replace quasi-identifier with the values with value that are less specific but semantically consistent. This process of replacement helps to arrange all quasi identifier values in a group that would be generalised to the entire group in the QID space. If at least two transaction in a group have distinct values in a certain column(i.e. one contains an item and other does not), then all information about that item in the current group is lost. Due to the high dimension of quasi identifier, it is likely that any generalization method would incur extremely high information loss. In order to maintain a high severity in generalization, records in the same bucket must be closed to each other so that generalizing the records would not lose much information. In spite of all these generalization do have various limitations such as:

- It fails on high dimension data due to its confined dimension.
- It causes too much of information loss due to uniform distributed assumption.

ii) BUCKETIZATION

Bucketization is technique of partitioning the tuples in T into buckets and then separating the sensitive attributes from non sensitive ones by randomly permuting the sensitive attribute value with each bucket. The sanitised data then consist of the bucket with the permuted sensitive values.

Bucketization first partitions tuples in the table into bucket and then separates the quasi identifier with sensitive attribute by randomly permuting the sensitive attribute value in each bucket. The anonymized data consist of a set of buckets with permuted sensitive attribute values. In spite of all these bucketization also have some limitation like:

- It does not prevent membership disclosure since it publishes the original values of Quasi Identifier in its original form so an adversary can find out the individual identity.
- It requires clear separation between Quasi Identifier and Sensitive attributes.
- It sometimes get confused between these two and hence brakes the correlation between them.

VI. PROPOSED SOLUTION FOR ANONYMIZATION

As for the privacy preservation a high dimensional database has become important in many ways. Database of online social network is so valuable that it must be preserved so that no confidential information should get leaked out. In this paper we have achieved a secure anonymization by partitioning the database in both horizontal as well as vertical form. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets, where in each bucket the values are randomly permuted to break the links between different column. This reduces the dimension of the data and does preserves better as compared to generalization and bucketization. The various strong point of dividing the database into horizontally and vertically are as follows:

- It protects privacy because it breaks the association between uncorrelated attribute which are infrequent and thus identifying.

- It group together QI and SA to preserve attribute correlation between them.
- Partitioning ensures that for each tuple there are generally multiple matching bucket.

Moving on to the second paradigm of our work i.e.; clustering which deals with data storage and its classification. The biggest problem of clustering in data mining is that it has to discover a new set of categories along with its privacy. On one hand we allow miners to discover useful knowledge from anonymized data On the other hand, we have to secure our data and set a limitation that how much the user can view the database, for this process various clustering models have been developed but here the main algorithm which have been used is NBDH.

i) CLUSTERING

In clustering Data obfuscation is the main method for securing of data, it breaks down the linkage between primitive data but the problem which lies here is about data loss. Data obfuscation blurs the individual characteristic whenever there is a breach of security and thus it becomes a pressing problem to preserve data utility as well as privacy. Here in this paper we propose a new technique of clustering named NBDH Neighbour based data hiding with K-anonymity (NBDH). The main principle of NBDH is to stabilize the nearest K neighbourhood. The benefits of NBDH are as follows:

- It preserves the data utility for clustering as well as privacy by sanitizing the nearest neighbour.
- It gives a good statistical approach for the data loss.
- It delivers a good set off between data utility and its trade off.
- It promotes swapping and data substitution based on its attributes.

5.1.1 DATA UTILITY

In current scenario, there are two strategies to achieve data utility for clustering, namely: distance-preserving perturbing and distribution-preserving perturbing. Initially the distances between data points are pointed out, and later data distribution is maintained. However distance preserving is over powerful to keep distances unchanged at the cost of weak security. As law et al [16] points out that distance-preserving is vulnerable to known Input–Output or Sample attacks. As shown in figure 2 given data point p, distance preserving solution needs to maintain relationship between p and rest of n-1 other data point.

On the other hand, distribution preserving solutions only maintain relations between p and the sketch of data subset. The mapping constraint of distance-preserving is far away stronger than that of distribution preserving. However, the weaker mapping constraint of distribution-preserving comes at cost of loss of individual characteristics.

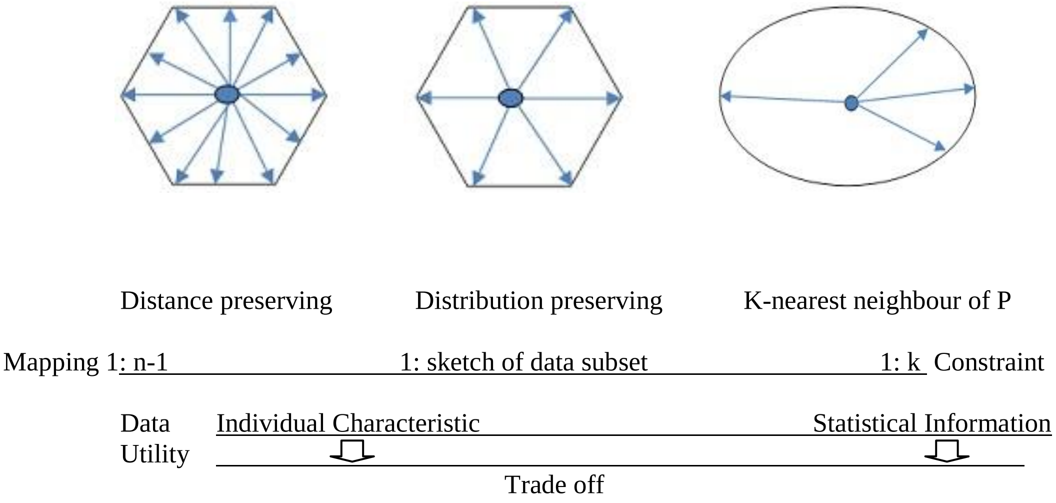


Figure 2 Comparison of obfuscation strategies.

5.2 PROBLEM STATEMENT AND ANALYSIS

Consider data set *D* with attribute set $A= \{A_1, A_d\}$ where attributes are numerical and assumed as sensitive attributes without loss of generality. Other notations are summarized in Table 4 and Let *d(p, q)* be the ordinary distance between two points p and *q*. Suppose obfuscation algorithm *f(.)* is applied on data set *D. f(p)* is the perturbed version of *p*. Function *l(p)* returns the identification of data point p. The stability of K- nearest neighbour of *D* can be defined as.

$$S(p) = 1 - \frac{|\{l(q)\,|\,q \in NNk\,(f(p)), p \in D, q \in D\}|}{|\{l(q)\,|\,q \in NNk\,(p), p \in D\}|} \quad (1)$$

$$S(D) = \frac{1}{n} \sum_{p \in D}^{s(p)} \quad (2)$$

The smaller value indicates the higher stability of K- nearest neighbour structure after the obfuscation *f(D*). The problem now here arises in generating a different version of *D* such that the possible privacy leakage is avoided while K- nearest neighbour structure is maintained.

VII. SOLUTION

In order to maintain the structure of K- nearest neighbour structure we first analyze the impact of p on $NN_k(p)$ by considering its attributes one by one. For example, in Fig. 3 $NN_k(p)$ exhibits more dispersed structure from p on A_i than that on A_j. This kind of dispersed/concentrated structures is the intrinsic structure characteristics of *NNk(p)*.The main idea is to differentiate these two types of attributes for each data

Symbols	Description
D	**Data set**
n, d	**Data Size and Dimension of D**
A_i	**The $_i$th attribute of data set**
p, q	**Data Points in D**
p^i	**Value of data point P on A_i**
$NN_k(p)$	**The set of k nearest neighbour of p**
$NN^i_k(p)$	**Set of data value in $NN_K(p)$ on A_i**

Table 4 Summary of notations

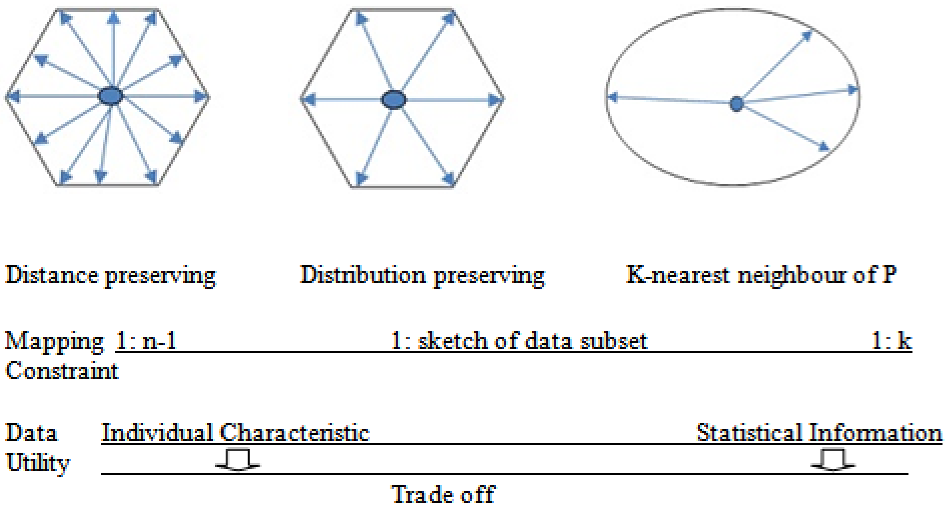


Figure 3 Distribution of K- Nearest Neighbour

6.1 NBDH: A MIXED MODE PERTURBING METHOD

In NBDH we have used Heuristic approach for data swapping which is simple and dynamic approach in itself. For P ∈ D, $A_i \in A$, S ⊂ D, Q ∈ S, $P_i \neq Q_i$ while making a swapping between P_i and Q_i the swapping cost will be calculated as SC_i (p, q, s) with the following formula

$$SC_i\,(p, q, s) = \frac{|\sum r \in s\,|ri - pi| - \sum r \in s\,|ri - qi||}{|S| \times |Pi - qi|}$$

From the definition, it can be deduced that the smaller the swap-ping cost is, the less distribution difference before and after swap-ping is. For p ∈ D, Ai ∈ A, if Ai is an CA of p, our swapping strategy substitutes pi with qi(pi≠ qi), where q is chosen within $NNk_{(p)}$ to minimize SC_i (p, q, $NNk_{(p)}$).

6.2 EXAMPLE OF NBDH

In this section we will try to explain our work with the help of graph and Tables, In Table 5 we have our original data set and Table 6 represents the k nearest neighbour relations of the data set (with k set to 3 and 4). Similarly setting parameter for k of NBDH to 3,the Dispersed attribute(DA) of neighbour and Concentrated attribute(CA) of neighbour of each point are differentiated as shown in Table 7, Table 8 demonstrates the perturbed data set by NBDH. In detail, the shadowed values are perturbed result via neighbouring statistical data substitution strategy and the rest is the result via neighbouring data swapping strategy. There exists a significant difference between data set in Table 5 and data set in Table 8. It provides a well protection to original data set. Besides, we can see that the nearest neighbour structures in Table 9 behaves similarly with that in Table 6.

ID	A1	A2	A3	A4	A5
T1	3.2	6.8	9.0	4.2	16.0
T2	2.7	6.0	7.9	5.6	11.6
T3	3.8	9.6	10.2	3.1	15.2
T4	8.6	8.8	8.0	3.8	12.2
T5	4.9	6.4	6.7	4.5	18.4
T6	5.2	5.6	7.0	7.7	10.5
T7	6.0	7.9	8.9	6.6	13.6
T8	1.2	5.2	7.8	5.0	18.0

Table 5 Original Dataset

ID	3 Nearest Neighbour data set	4 Nearest Neighbour data set
T1	t1 {t3, t5, t8}	t1 {t3, t5, t7, t8}
T2	t1 {t1, t6, t7}	t1 {t1, t3, t6, t7}
T3	t1 {t1, t5, t7}	t1 {t1, t4, t5, t7}
T4	t1 {t3, t6, t7}	t1 {t2, t3, t6, t7}
T5	t1 {t1, t7, t8}	t1 {t1,t3, t7, t8}
T6	t1 {t2, t4, t7}	t1 {t1, t2, t4, t7}
T7	t1 {t2, t4, t6}	t1 {t1, t2, t4, t6}
T8	t1 {t1, t3, t5}	t1 {t1, t2, t3, t5}

Table 6 K- Nearest Neighbour Relation

ID	**DA**	**CA**
T1	{A1, A4, A5}	{ A2, A3}
T2	{ A1, A3, A4}	{ A2,A5}
T3	{A4}	{ A1, A2, A3, A5}
T4	{A4}	{ A1, A2, A3, A5}
T5	{A2,A3}	{ A1, A4, A5}
T6	{ A1, A2, A3, A4}	{A4}
T7	{ A2, A3, A5}	{ A1, A4}
T8	{ A1, A3, A4, A5}	{A2}

Table 7 DA and CA

Table 5, 6, 7 representing Pre-treated Data Set

ID	A1	A2	A3	A4	A5
T1	3.3	6.4	7.8	4.2	17.2
T2	4.8	6.8	8.3	6.2	13.6
T3	4.9	6.4	6.7	5.1	16.0
T4	3.8	7.9	8.9	5.8	13.6
T5	3.2	6.6	8.6	5.0	16.0
T6	5.8	7.6	8.3	3.8	12.5
T7	5.2	6.8	7.6	5.6	11.4
T8	3.9	9.6	8.6	3.9	16.5

Table 8 Perturbed Data Set

ID	**3 Nearest Neighbour data set**	**4 Nearest Neighbour data set**
T1	t1 {t3, t5, t8}	t1 {t3, t4, t5, t8}
T2	t1 {t4, t6, t7}	t1 {t3, t4, t6, t7}
T3	t1 {t1, t2, t5}	t1 {t1, t2, t4, t5}
T4	t1 {t2, t5, t6}	t1 {t2, t5, t6, t7}
T5	t1 {t1, t3, t4}	t1 {t1,t2, t3, t4}
T6	t1 {t2, t4, t7}	t1 {t2, t3, t4, t7}
T7	t1 {t2, t4, t6}	t1 {t2, t3, t4, t6}
T8	t1 {t1, t3, t5}	t1 {t1, t2, t4, t5}

Table 9 Perturbed K-Nearest Neighbour Relation Data Set

Table 8, 9 representing Perturbed Data Set with NBDH

6.3 ADVANTAGES OF NBDH

- NBDH has totally different rotation angels from RBT [17], parameter k in NBDH cannot be inferred merely by grasping several pairs of points.
- Attackers will be unable to determine the exact range of the nearest neighbour set and the distinction criterion of DA and CA. These provide the core foundation of security for NBDH.

- For each data point, the division of its attributes is an irreversible process. Attackers cannot re-construct original attribute division by the public perturbed data. Therefore, even if attackers have grabbed value of k, they fail to make further inference on exact type of attributes of data points.
- NBDH provides a full guarantee of security from adversary as the point of rotation is very complex as in the pattern are always different.
- NBDH not simply perturbs data set merely with data swapping method. It adopts a mixed mode of data swapping and also it adopts statistical data substitution strategy.

6.4 ADVANTAGE OF PROPOSED TECHNIQUE

This paper explores all the possibility of cleaning and securing a social network to prevent inference of social network data and then examines the effectiveness of those approaches on a real-world data set. In order to protect privacy, we have worked on the both paradigm Anonymization as well as for clustering, we deleted some of the information from a user's profile and manipulated some links between friends. We also examined the effects of generalizing detail values to more generic values. We have proposed a new technique for securing of the information and there after we have worked on the database clustering method which stores the data sets in a distinguish form and show the new technique so that we can calculate the generalised form of the lost information in databases.

VIII. CALCULATION OF THE INFORMATION LOSS

Memory consumption and information loss is a very important performance metric, to know the efficiency of a particular algorithm. Memory consumption can be defined as total number of memory used in bytes for RAM and ROM. If RAM and ROM value is more, then overhead is also more. similarly to know the information loss we have taken some data into consideration and tried to show the data loss in three modes which are Structural loss of information, Generalised information loss and lastly total information loss. Different values have been taken for consideration along with different valuation formulas Figure 4 shows the Generalised Information Loss between clusters along with Figure 5 which shows Structural Information Loss and at last Figure 6 depicts the Total Information Loss, For calculation of information loss we have considered different clusters along with their values

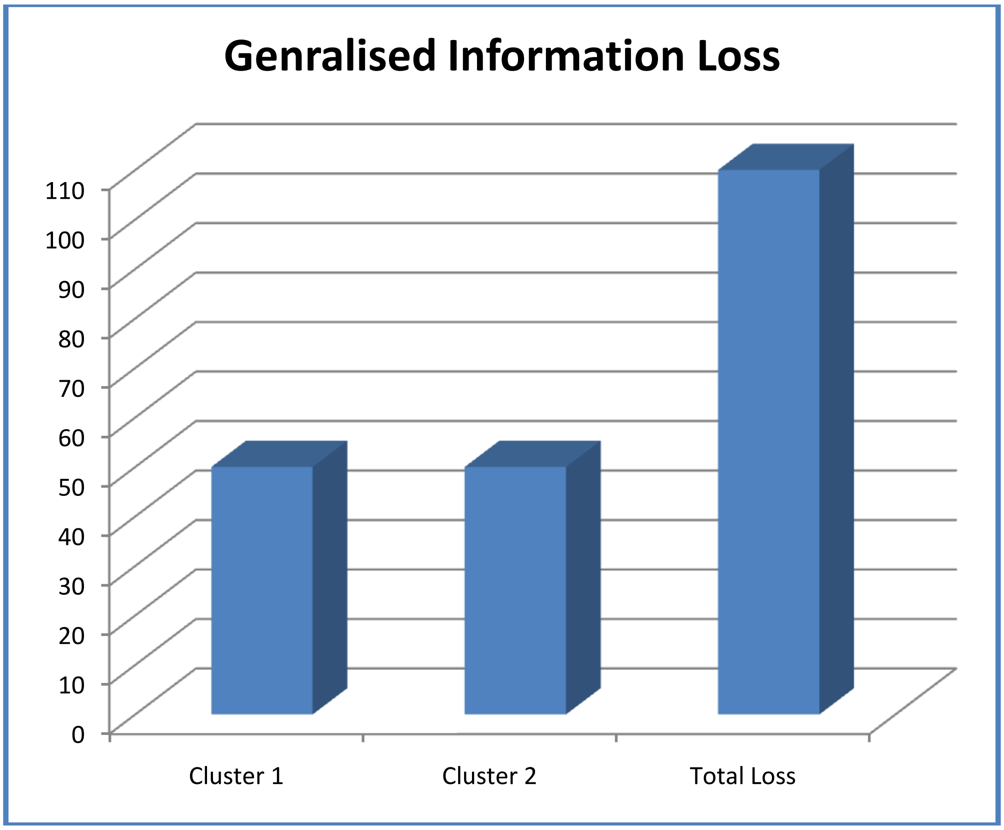


Figure 4 Generalised Information Loss

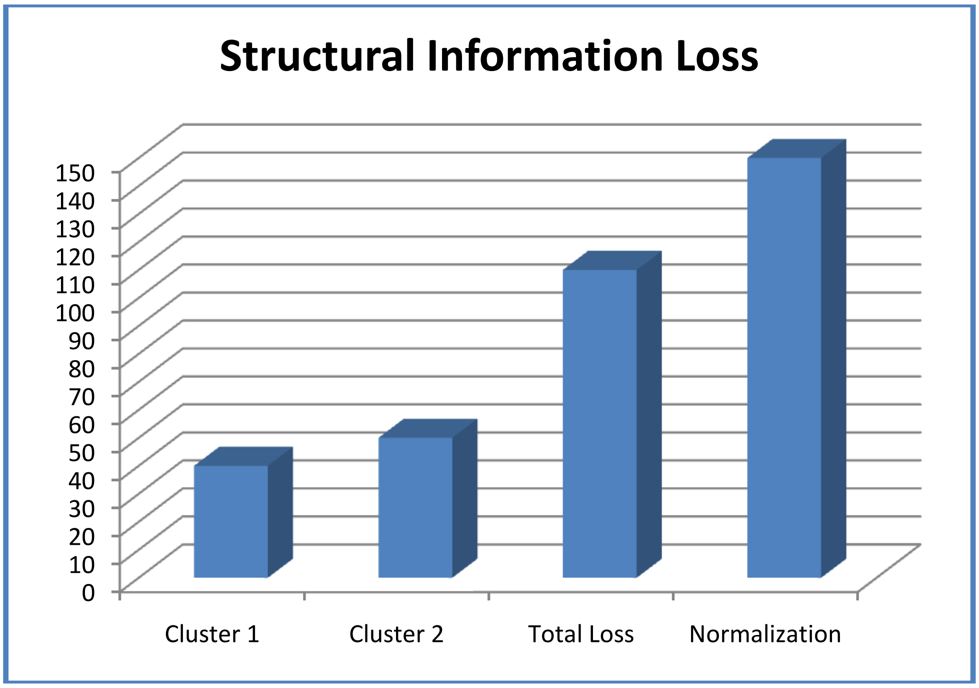


Figure 5 Structural Information Loss

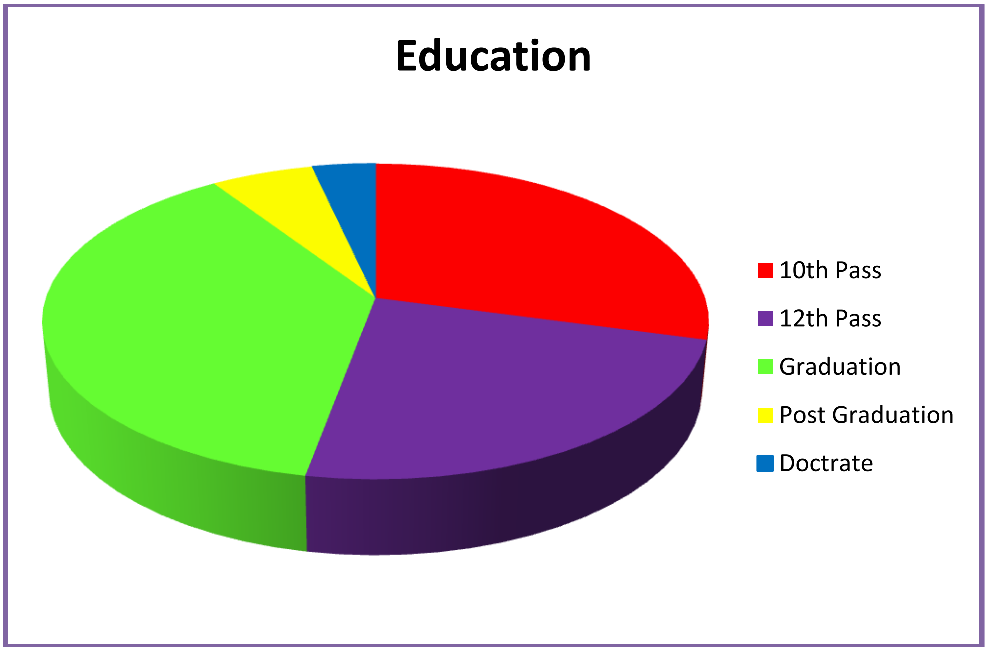


Figure 6 Total Information Loss

IX. CONCLUSION

Firstly, understanding the concept of K-anonymity with clustering is done. Successful implementation of Enhanced K-anonymity is done where a new approach of partitioning has emerged which divides the database into Vertical and Horizontal partitioning. A new algorithm NBDH has been proposed to make one of the clustering standard available in market, It preserves clustering quality by maintaining the stability of nearest neighbourhoods. A mixed mode of data swapping and substitution perturbing methods is developed for attributes of different types and it helps to explore the path to implement such an algorithm with the combination of K-Anonymity and Clustering. Mainly, our concept is the combination of K-Anonymity and Clustering.

This is a Key dependent algorithm which has control over the QA and SA. We addressed various issues related to private information leakage in social networks. In addition, the effect of removing details and links in preventing sensitive information leakage is also explored. We have worked on both areas of social networking i.e. Anonymization as well as Clustering and at the end it was found that by removing only details the accuracy of local classifiers, which give us the maximum accuracy that we were able to achieve through any combination of classifiers was reduced. We also assumed full use of the graph information when deciding which details to hide. Useful research could be done on how individuals with limited access to the network could pick which details to hide.

FUTURE WORK

Future work could be conducted in identifying key nodes of the graph structure to see if removing or altering these nodes can decrease information leakage. Several directions for future research exist, including adapting NBDH to high dimensional data sets, as well as incrementally obfuscating data sets.

REFERENCES

[1]. http://en.wikipedia.org/wiki/Facebook.

[2]. http://en.wikipedia.org/wiki/Twitter.

[3]. http://en.wikipedia.org/wiki/Linked_in.

[4]. http://en.wikipedia.org/wiki/Vkontakte.

[5]. Gross, R. and Acquisti, A. 2005. Information Revelation and Privacy in Online Social Networking Sites (The Facebook Case).[online]. p. 2. Available at :http://www.heinz.cmu.edu/~acquisti/papers/privacy-facebook-gross-acquisti.pdf

[6]. http://articles.latimes.com/2012/feb/14/business/la-fi-tn-twitter-contacts-20120214

[7]. http://www.foxnews.com/scitech/2012/02/16/twitter-admits-peeking-at-address-books-announces-privacy-improvements/

[8]. http://en.wikipedia.org/wiki/AOL_search_data_leak

[9]. M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava, “Anonymizing Social Networks,” Technical Report 07-19, Univ. of Massachusetts Amherst, 2007.

[10]. K. Liu and E. Terzi, “Towards Identity Anonymization on Graphs,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’08), pp. 93-106, 2008.

[11]. E. Zheleva and L. Getoor, “To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private user Profiles,” Technical Report CS-TR-4926, Univ. of Maryland. 2008.

[12]. J. He, W. Chu, and V. Liu, “Inferring Privacy Information from Social Networks,” Proc. Intelligence and Security Informatics, 2006.

[13]. Latanya Sweeney. k- anonymity: a model for protecting privacy. International journal of uncertainty, Fuzziness and Knowledge-Based Systems,10(5):557-570,2002.

[14]. A. Machanavajjhala, D.Kifer, J.Gehrke and M. Venkitasubramanium. "I-Diversity: Privacy beyond k-anonymity". In ICDE,2006.

[15]. Kun Liu, Chris Giannella, Hillol Kargupta, An attacker’s view of distance preserving maps for privacy preserving data mining, in: PKDD 2006, Berlin, Germany, September 18–22, 2006, pp. 297–308.

[16]. S.R.M. Oliveira, O.R. Zaiane, Achieving privacy preservation when sharing data for clustering, in: SDM’04,Toronto, Ontario, Canada, 2004, pp. 67–82.