Privacy Preservation using T-Closeness with Numerical Attributes

Sheshang D. Degadwala
HeadCom. Eng. Department
Sigma Inst. of Eng.
Vadodara, Gujarat

Arpana D. Mahajan
Asst. Prof. Com. Eng. Department
Sigma Inst. of Eng.
Vadodara, Gujarat

Dhairya J. Vyas
Asst. Prof. E&C Department
Sigma Inst. of Eng.
Vadodara, Gujarat

Abstract*:* Data mining is a process that is used to retrieve the knowledgeable data from the large dataset. Information imparting around two associations will be basic done a large number requisition zones. As people are uploading their personal data over the internet, however the data collection and data distribution may lead to disclosure of their privacy. So, preserving the privacy of the sensitive data is the challenging task in data mining. Many organizations or hospitals are analyzing the medical data to predict the disease or symptoms of disease. So, before sharing data to other organization need to protect the patient personal data and for that need privacy preservation. In the recent year's privacy preserving data mining has being received a large amount of attention in the research area. To achieve the expected goal various methods have been proposed. In this paper, to achieve this goal a pre-processing technique i.e. k-means clustering along with anonymization technique i.e. k-anonymization and t-closeness and done analysis which techniques achieves more information gain.

Keywords:*k-anonymization, t-closeness, k-means clustering*

*****

I. Introduction

Data mining aims to retrieve a useful or knowledgeable data from the large data repository. But privacy preservation means publishing the knowledgeable data along with hiding/protecting the sensitive data (personal data). Concerning illustration the aggregate utilization of information mining, vast information from claiming private majority of the data (data) would recurrently gathered Furthermore broke down. Such information hold numerous shopping habits, restorative history, criminal records, kudos records and so forth throughout this way, observing and stock arrangement of all instrumentation may be etc [1]. And these data are the important asset to the governments and business organization for conclusion making process and to deliver social benefits such as health research, nationwide security etc. This is mostly concern through government agencies, insurance companies, hospitals and others that have information they would like to issue to researchers.Every record takes a number of attributes which can be separated into 3 categories: I. Explicit identifiers which can identify people. II. Quasi identifier features whose values when taken can simply recognize entity's characteristics. III. Sensitive attributes which are considered as private and need not be revealed [2]. Different anonymization methods have been explored to guard the sensitive data of the people. Re-identification of the data does not deliver anonymity, as the out data too contain quasi identifiers can be used for re-identifying the information of people, thus dripping that data which is not planned to be revealed data leakage happens by synchronization of quasi identifiers and outsideinformation. Thus, the main aim is to bind the expose risk towards an acceptable level. So, it can be reached by anonymizing the data previously discharging it.

Distribution of anonymization techniques: data anonymization is defined as the use of one or more techniques designed to make it difficult to identify the sensitive/personal information from the stored data set. Fig. 1.Showstypes of anonymization methods.

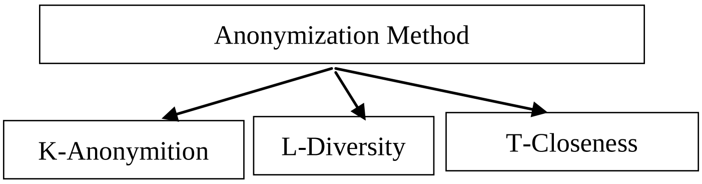


Fig. 1. Classification of anonymization techniques [2]

II. Methodology

K-anonymity: Once liberating micro information for Scrutinize intention, particular case wishes to bind uncover dangers to a palatable stage if expanding information convenience [1]. In the heading about bound uncover risk, there need been exhibited k-anonymity security necessity, that needs each distinct passage over a anonym zed table with make unintelligible through in any event k-1 additional entrance inner of the information set, for admiration to quasi-identifier qualities. On the way with accomplish the k-anonymity state, they utilized commonly generalization likewise concealment to the information anonymization. Anonymization may be the system that dispenses with Overall substitute's uniqueness information starting with a record.

Fig 2.Displays micro data table. Fig. 3.Displays voter's registration information. Fig. 4.Displays a sample of 2-anonymous generalization designed for Fig. 2. By the voter registration list, an opponent will individual conclude that Jim might be the person involute in the primary two tuples of Figure 4, else consistently, the actual disease of Jim stays exposed individual by probability 50%.

ID	Age(yr)	Gender	Zip Code	Disease
1	14	F	85651	Cough
2	13	F	85654	Cough
3	12	F	85957	Toothache
4	29	M	85959	Headache

Fig. 2. Micro data [1]

ID	Name	Age(yr)	Gender	Zip code
1	Jim	14	F	85651
2	Jaya	13	F	85654
3	Timi	12	F	85957
4	Luv	29	M	85959

Fig. 3. Voter Registration List [1]

ID	Age(yr)	Gender	Zip Code	Disease
1	1*	F	856**	Cough
2	1*	F	856**	Cough
3	1*	F	859**	Toothache
4	2*	*	859**	Headache

Fig. 4. A-2 Anonymous Table [1]

The micro data may contain 4 categories of attributes that is explicit identifiers, quasi identifiers (QIs), sensitive attributes (SAs), non-sensitive attributes [11]. K-anonymity protects in contradiction of individuality expose; it does not offer enough guard in contradiction of attribute expose [7]. Dual eruptions were recognized homogeneity attack & background knowledge attack. Linking attack is performed by taking exterior tables holding the individualities of persons, & some otherwise entirely of the public attributes.

L-diversity: It efforts to guard in contrast to attribute disclosure by taking 'l' discrete standards aimed at individually sensitive feature in a set of queues in mixture along with further features [5]. L-diversity stands a real-world, easy to know, & reports the inadequacies of k-anonymity by background knowledge & homogeneity attacks [12]. L-Diversity dose not needs information of the complete distribution of the sensitive & non-sensitive attributes. It doesn't need the information producer to take as considerable data as the enemy. The limit guards against well-informed enemies; the bigger the value of l, further the data is required to rule out probable standards of the sensitive attributes [12]. Different enemies can have dissimilar background information leads to dissimilar inferences. L-Diversity at the same time guards in contrast to all of them without the necessity for inspection which implications can be made with which stages of background knowledge. It has 3 kinds that is recursive l-diversity, entropy based diversity & distinct l-diversity [5].

T-closeness: Those comparability population will be assumed on detract t-closeness accepting that those separation around spreading of a delicate characteristic in the class & spreading of the delicate characteristic in the whole table may be not additional over an edge 't'. Table may be assumed should need t-closeness if wholly comparability classes detract t-closeness [7]. The‘t’ parameter done t-closeness empowers you quit offering on that one should tradeoff the middle of utility and security. T-closeness anonymity procedure sustains subsequent property: the detachment among the spreading to informative attribute from thegroup&spared sensitive featurefor an whole data set and it’s not extra than the threshold ‘t’. t-closeness method too is vulnerable to skewness & similarity attacks. Writers hastoo cited that if threshold‘t’data will be enlarged, then diminishes t-closeness possessions & weakens the connection among key attributes and sensitive attributes [5].Thus, decided that defending data secrecy is a significant matter while gathering micro data. The k-anonymition can be struggle relatives attacks, L-diversity know how to struggle in contrast to standardized attacks, t-closeness resolve bright with lessen the data damage.

III. Proposed Method

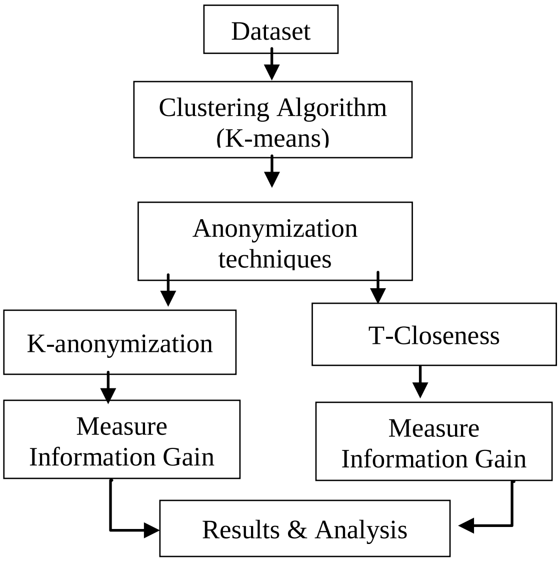


Fig 5: Proposed Block Diagram

K-meansgrouping: Grouping is the procedure of dividing a cluster of data factsaddicted to a minorno. of groups.This method commonly used in the automatically partition a data set into k groups.

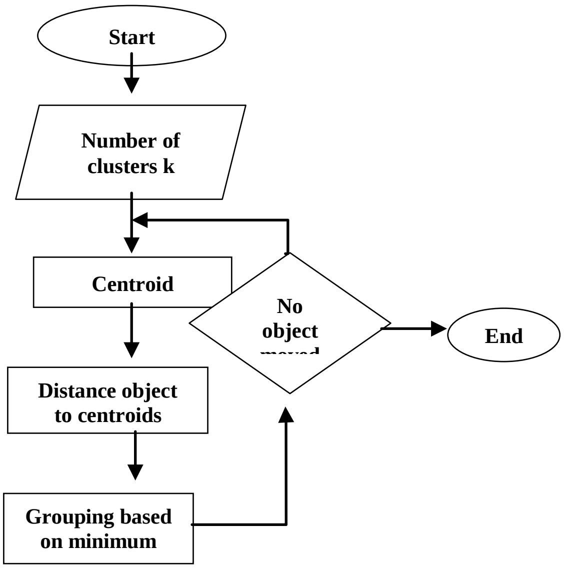


Fig 6: Flow diagram of k-means clustering

K-anonymization: AssumedPeople’s structure data, make theissue of the data bytechnicalpromises.For thedifferent personsthat are whofor subjects of the information can't make re-identified same time those information stay practically suitable.

k-anonymitionmethod for the k's value:

1. Generalization: Level of detail in dataset is reduced. Make less informative or replace the attribute value with its more generalized value.

Example: Replacing the age class with a particular range i.e. if age is 22 then it can be replaced with range of 20-24.

2. Suppression: The value of attribute is completely removed.

Example: In this the whole value, will be replaced with some symbol such as "*" or "?".

t-closeness: Asimilaritycls is supposed to be *t*-closeness then the spaceamong themsupply of a sensitive attribute tothatclsforspreading of attribute to the entire table is num.additionalwiththe threshold value *t*.

EMD (Earth mover separation) obliges that those separation between those two probabilistic circulations will be subordinate upon those ground distances around the values of a trait. The primary point for EMD will be that it has the ability should catch those semantic separation between values. The separation between two probabilistic circulations might be measured utilizing world Mover's separation (EMD). EMD obliges that the separation the middle of the two probabilistic circulations will a chance to be indigent upon those ground distances around those values about a trait. To numerical attributes, the common request might make used to measure such separation. Give X=(x1, x2,…. ,xn), Y=(y1, y2,…. , yn) make the provided for two circulations Also d_ij make the ground separation the middle of those component i about X What's more component j for Y. Whether the component 1 need an additional measure of x1-y1, then the measure about y1- x if make transported starting with different components of the component 1. Also also, component 1 is transported from component 2. After that component 1 is fulfilled What's more component 2 need an additional measure of (x1-y1) + (x2-y2).The over depicted transform proceeds until component m is fulfilled and Y may be arrived at. Lesvos rk = xk-yk, (k=1, 2. N)At that point the separation the middle of X what's more Y could make computed as.

D[X Y]= $\frac{1}{n-1}$("$|r_1|$+$|r_1 + r_2|$+ ………+ $|r_1 + r_2 + r_{n-1}|$")

n=k

= $\frac{1}{n-1}\sum|\sum r_j|$, k=1, j=1 [1]

For categorical several distances are proposed according to the type of categorical attribute under consideration.

D [X Y] = $\frac{1}{2}\sum|x_i - y_i|$, i=1 [2]

q=[2000,3000,4000,5000,6000,7000,8000,9000,10000,11000]p1=[2000,3000,4000,5000]

p2=[7000,8000,9000,10000]

D(q,p1)=0.3889

D(q,p2)=0.211

Experimental Parameters:

Information Gain(S) defined as

Info Gain (v) = I (Rs)-$\sum_c \frac{|Rs|}{|Rv|}$ I(Rc), [3]

Where s is for child (s)

Rv is maximum number of records in particular class

Where I (Ry) is the entropy

I (Ry) = -$\sum_{cls} \frac{freq\,(Ry,cls)}{|Ry|} \times log_2\, freq\, \frac{freq\,(Ry,cls)}{|Ry|}$, [4]

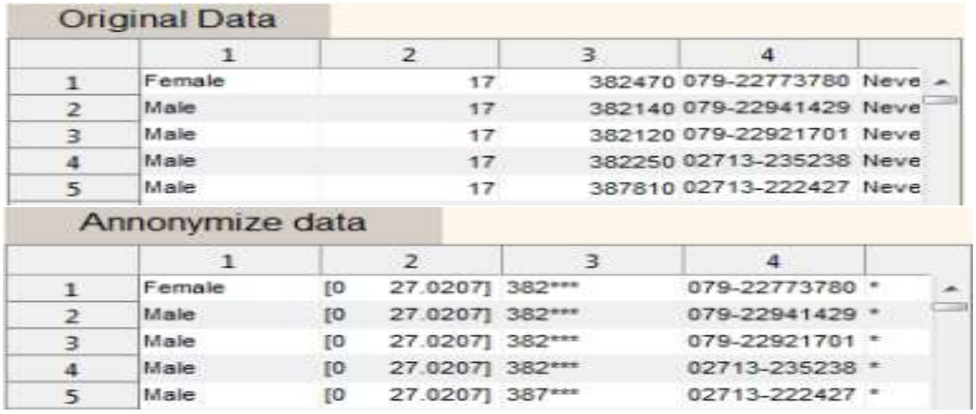
Where Rx is total number of records and $freq(Ry, cls)$ is the no. of data records in Ry having class Anonymity loss is defined as,

AnonyLoss (s)= $avg\,\{A(VID_j) - Av(VID_j)\}$, [5]

Where, VID is the total number of records and Av is number of records of particular class

Score(v)= $\begin{cases} \frac{InfoGain\ (v)}{AnonyLoss\ (v)} & if\ AnonyLoss\ \neq\ 0 \\ InfoGain(v) & otherwise \end{cases}$, [6]

IV. Experimental And Analysis



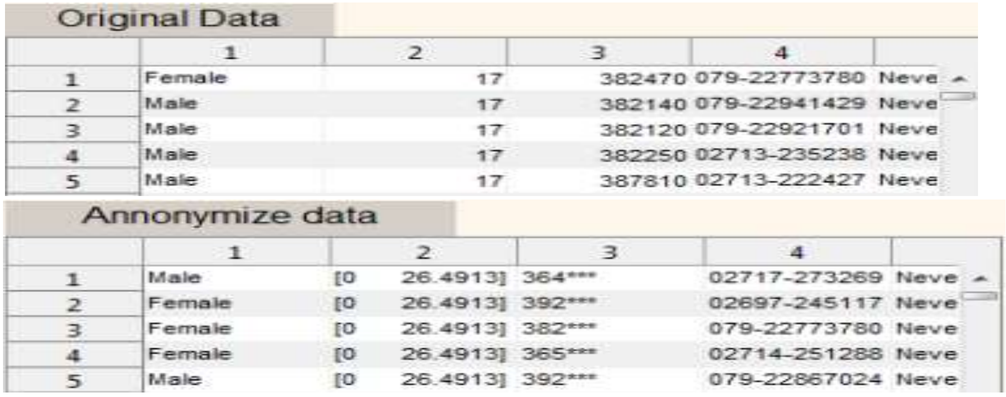
Original Data

	1	2	3	4	
1	Female	17	382470	079-22773780	Neve
2	Male	17	382140	079-22941429	Neve
3	Male	17	382120	079-22921701	Neve
4	Male	17	382250	02713-235238	Neve
5	Male	17	387810	02713-222427	Neve

Annonymize data

	1	2	3	4	
1	Female	[0 27.0207]	382***	079-22773780	*
2	Male	[0 27.0207]	382***	079-22941429	*
3	Male	[0 27.0207]	382***	079-22921701	*
4	Male	[0 27.0207]	382***	02713-235238	*
5	Male	[0 27.0207]	387***	02713-222427	*

Fig 7: k-anonymization



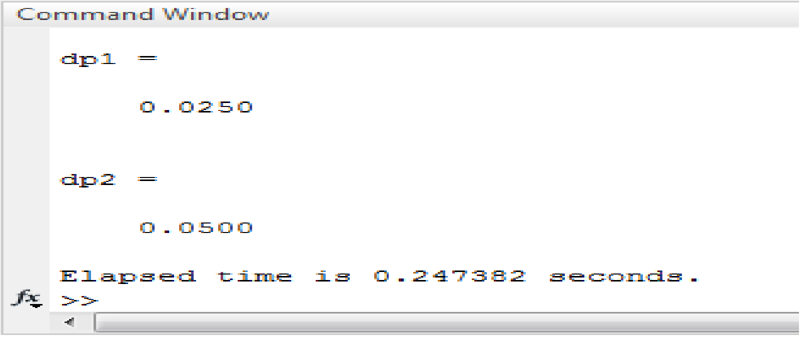
Original Data

	1	2	3	4	
1	Female	17	382470	079-22773780	Neve
2	Male	17	382140	079-22941429	Neve
3	Male	17	382120	079-22921701	Neve
4	Male	17	382250	02713-235238	Neve
5	Male	17	387810	02713-222427	Neve

Annonymize data

	1	2	3	4	
1	Male	[0 26.4913]	364***	02717-273269	Neve
2	Female	[0 26.4913]	392***	02697-245117	Neve
3	Female	[0 26.4913]	382***	079-22773780	Neve
4	Female	[0 26.4913]	365***	02714-251288	Neve
5	Male	[0 26.4913]	392***	079-22867024	Neve

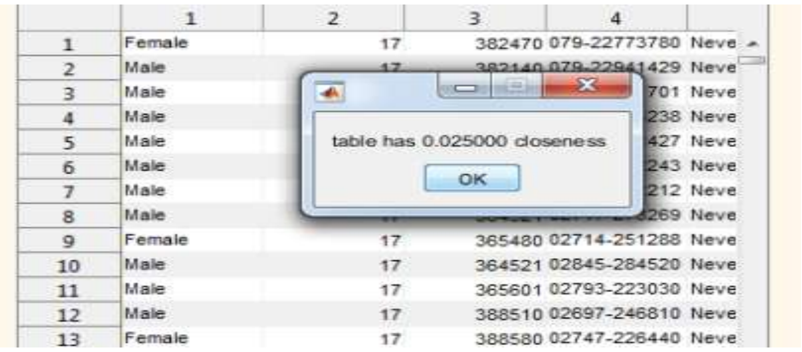
Fig 8: t-Closeness



Command Window

```
dp1 =

     0.0250


dp2 =

     0.0500

Elapsed time is 0.247382 seconds.
>>
```

Fig 9: EMD Calculation



	1	2	3	4	
1	Female	17	382470	079-22773780	Neve
2	Male	17	382140	079-22941429	Neve
3	Male			701	Neve
4	Male			238	Neve
5	Male			427	Neve
6	Male			243	Neve
7	Male			212	Neve
8	Male			269	Neve
9	Female	17	365480	02714-251288	Neve
10	Male	17	364521	02845-284520	Neve
11	Male	17	365601	02793-223030	Neve
12	Male	17	388510	02697-246810	Neve
13	Female	17	388580	02747-226440	Neve

table has 0.025000 closeness

Fig 10: table Closeness 0.025000

Table 1: Parameters

Methods	Information Gain	Anonymity Loss	Score
k-Anonymization	0.75229	1917	0.00040195
t-Closeness	0.82255	1586	0.00051253

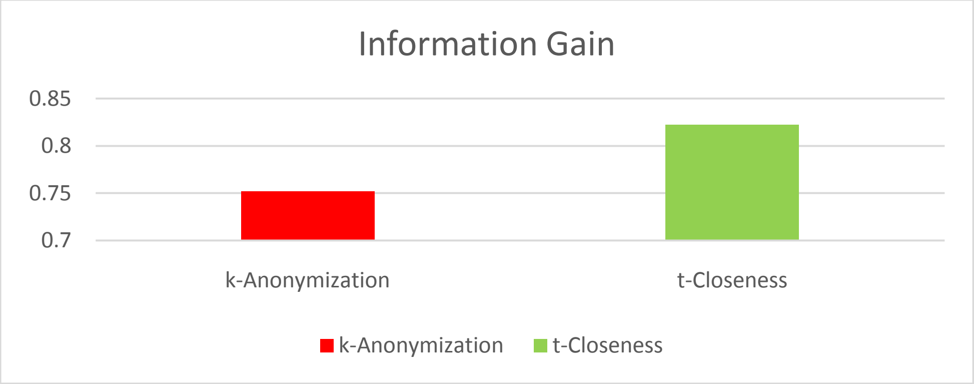


Fig 11: Information Gain

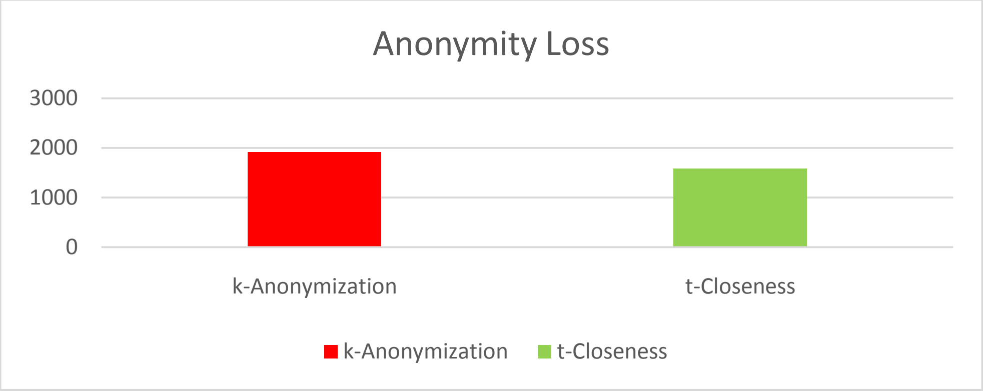


Fig 12: Anonymity Loss

V. Conclusion

Confidentiality is the major disquiet to protect the sensitive data. People are aware of their sensitive infoand they don’t want to share to anybody. The anonymization algorithm are used reduced information loss and increase the privacy protection. . t-closeness, isneedsforsharing sensitive attribute in any sameness class is near todelivery for attribute contain overall table data (i.e., the distance to a2-allocationsnot to be excite threshold t). It’s provide better privacy, security and encryption to the system.In privacy preservation comparing the information gain and anonymity loss for k-anonymization and t-Closeness(numerical attribute) has been done on the quasi-identifier attribute and conclude that the information gain will be more and anonymity loss will be less in t-Closeness compared to k-anonymization.

References

[1] S. Lohiya and L. Ragha, "Privacy Preserving in Data Mining Using Hybrid Approach," Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on, Mathura, 2012, pp. 743-746.

[2] Dilpreet Kaur Arora, Divya Bansal and SanjeevSofat, “Comparative Analysis of Anonymization Techniques,” International Journal of Electronic and Electrical Engineering, ISSN 0974-2174 Volume 7, Number 8 (2014), pp. 773-778.

[3] M. Sharma, A. Chaudhary, M. Mathuria, S. Chaudhary and S. Kumar, "An efficient approach for privacy preserving in data mining," Signal Propagation and Computer Technology (ICSPCT), 2014 International Conference on, Ajmer, 2014, pp. 244-249.

[4] P. Canbay and H. Sever, "The Effect of Clustering on Data Privacy," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, 2015, pp. 277-282.

[5] K. S. Banu, V. Santhi and B. K. Tripathy, "Non-cryptographic security to data: Distortion based anonymization techniques," Advances in Engineering and Technology (ICAET), 2014 International Conference on, Nagapattinam, 2014, pp. 1-5.

[6] R. B. Ghate and R. Ingle, "Clustering based Anonymization for privacy preservation," Pervasive Computing (ICPC), 2015 International Conference on, Pune, 2015, pp. 1-3.

[7] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l Diversity," 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, 2007, pp. 106-115.

[8] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez and S. Martínez, "t-closeness through microaggregation: Strict privacy with enhanced utility preservation," 2016 IEEE 32nd International Conference on Data Engineering (ICDE), Helsinki, 2016, pp. 1464-1465.

[9] A. V. Vashkevich and V. G. Zhukov, "Privacy-preserving clustering using C-means," Control and Communications (SIBCON), 2015 International Siberian Conference on, Omsk, 2015, pp. 1-4.

[10] V. Baby and N. S. Chandra, "Distributed threshold k-means clustering for privacy preserving data mining," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 2016, pp. 2286-2289.

[11] J. J. Panackal, A. S. Pillai and V. N. Krishnachandran, "Disclosure risk of individuals: A k-anonymity study on health care data related to Indian population," Data Science & Engineering (ICDSE), 2014 International Conference on, Kochi, 2014, pp. 200-205.

[12] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 2006, pp. 24-24.

[13] B. C. M. Fung, K. Wang and P. S. Yu, "Top-down specialization for information and privacy preservation," 21st International Conference on Data Engineering (ICDE'05), 2005, pp. 205-216.

[14] Nirav.U.patel, Vaishali.r.patel,“anonymization of social networks for reducing Communication complexity and information Loss by sequential clustering,” International Journal of Modern Trends in Engineering and Research (IJMTER) Volume 02, Issue 05, [May – 2015] ISSN (Online):2349–9745 ; ISSN (Print):2393-8161