# K-Anonymization Approach FOR Privacy Preserving IN Data Mining

**Vimalkumar B. Vaghela**

**Abstract:** Data is collected and processed using diverse sources and tools that lead to privacy issues. Till now randomization, k-anonymization model, l diversity, t closeness, cryptography and many more techniques have been used to preserve the privacy of an individual. But each and every technique have their own demerits i.e. Information Loss, Privacy breached, Low Data Utility. Among all these approach, k anonymization approach one of the mostly used anonymization based approach. However, this approach suffers from the issue of information loss. So, it is challenging task for data miner to mine data. The paper focus to decrease information loss using the 2 level k anonymization approach and also preserve privacy as compared to the traditional approach. The main aim of this approach is to decrease data loss and no compromise with privacy.

**Index Terms:** Anonymization, Data Mining, Generalization, Information Loss, Information Santization, Privacy Preserving, Suppression

## 1 INTRODUCTION

Privacy-Preserving after heard this word first thought comes to mind is that what kind of privacy will be preserved. Privacy preservation is required when some organization gives their user's data to another organization for a specific purpose. For example, Canatics is a company in Canada. It works is to find fraud from insurance claimers for that canatics have to collect all the data of insurance companies for fraud detection but because of it some private or sensitive information of insurance holders will be released in front of canatics. As a result, protecting the private information of an individual becomes a prime research issue in privacy preserving data mining [2]. Here k-Anonymization will help. k-Anonymization Approach is The concept of k-anonymity was first introduced by Latanya Sweeney and Pierangela Samarati in a paper published a publication of information is supposed to have the[ k-anonymity possessions if the info for each individual contained in the publication cannot be eminent from at least k-1 persons whose info also appear in the publication.

## 2 APPROACH AND FORMULAS

### 2.1 k-Anonymization Approach

This approach is better in cost, time complexity as compared to other approaches of privacy-preserving but, Information Loss and data utility this both are major issues of this approach. Mainly data set attributes classified in three types of attributes identifiers attributes (name, ssid number), quasi identifier attributes QI (age, zip code, gender), sensitive attributes SA (disease, salary). It can be further classified in numerical and categorical attributes. After classification generalization and suppression process will be applied on dataset [6]. Generalization as per name it will replace general value of that attribute in place of original value for example have quasi attribute Age and values are 31, 32, 33, 34, 35 and 36, then they can be represented as (31-36). In other hand suppression hides unique digits of the k records with same example 31, 32, 33, 34, 35 and 36, then they can be represented as 3*.

• *Dr. Vimalkumar B. Vaghela is currently working as an Assistant Professor in Computer Engineering Department at L. D. College of Engineering, Ahmedabad, Gujarat, India E-mail: vimalvaghela@gmail.com*

Here data loss is higher in suppression than generalization. To explain information loss because of generalization and suppression is described by example of medical dataset this medical organization wants to share their patient's data with research institute for research and mining reasons. Medical dataset is given in Table I. After applied Generalization and Suppression information get loss for example in table II original value of age replaced with range of age values, in attribute gender 'M' or 'F' values replaced with 'P' because of generalization and in zip code last digits by which patient record can be identified those digits are also hide by star so ultimately privacy is preserved but information got loss so, authors Pawan Baladhare and Devesh Jinwala[1] discussed this information loss problem and also introduced a novel approach which decrease a IL at some extent. Formulas to Calculate Information LossLet dataset D consists of a set of r tuples with n numeric and c categorical quasi identifier attribute. Let $\gamma = \{\sigma 1, \sigma 2, \sigma 3, \ldots, \sigma p\}$ be a dividing of $\gamma$, where $\sigma$ characterize a cluster Let ζ be the taxonomy tree as shown in Figs. 1 and 2 with respect to Table II. The taxonomy tree is used to generalize the value of each categorical and numerical attribute. The info loss (ILn) for the n numerical attributes of a dataset D via generalization and suppression is calculated as follows: Let $Ri_{max,}\ Ri_{min}$ be the maximum and minimum value of the tuples in a cluster . Let $Ti_{max,}\ Ti_{min}$ be the maximum and minimum value of the tuples in a dataset D.

$$IL_n = |\gamma| \sum_{i=1}^{n} \frac{Ri_{max} - Ri_{min}}{Ti_{max} - Ti_{min}} \quad (1)$$

***TABLE I***

*SUBDATABASE FROM HOSPITAL'S ORIGINAL DATASET*

| *Identifiers* | *Quasi-identifier* | | | *Sensitive* |
|---|---|---|---|---|
| *NAME* | *AGE* | *GENDER* | *ZIP CODE* | *DISEASE* |
| *AVS* | *22* | *M* | *132011* | *FLU* |
| *FVD* | *21* | *F* | *132150* | *HIV* |
| *RZG* | *15* | *F* | *132012* | *CANCER* |
| *TAH* | *20* | *M* | *132012* | *DIABETES* |
| *YDK* | *30* | *F* | *132150* | *CANCER* |
| *EDT* | *19* | *M* | *132011* | *HIGH BP* |
| *RNE* | *17* | *M* | *132050* | *DIABETES* |
| *WGE* | *41* | *F* | *132012* | *CANCER* |
| *GDA* | *27* | *M* | *132012* | *HIV* |
| *BEA* | *45* | *M* | *132150* | *CANCER* |
| *CKL* | *49* | *F* | *132011* | *FLU* |
| *XHI* | *43* | *M* | *133150* | *HIV* |
| *BFX* | *49* | *M* | *133012* | *CANCER* |
| *BSN* | *50* | *F* | *132012* | *DIABETES* |

| | | | | |
|---|---|---|---|---|
| *JSM* | *45* | *M* | *132153* | *CANCER* |
| *OAE* | *42* | *F* | *132453* | *HIGH BP* |

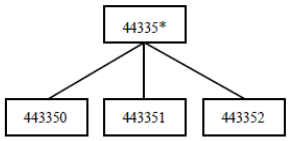


***Fig.1****. Taxonomy tree of zip code.*

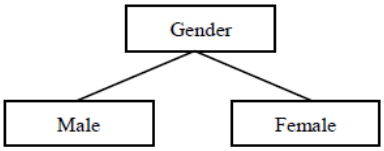


***Fig.2****. Taxonomy tree of gender.*

Similarly, the information loss (ILc) for the c categorical attribute is calculated as follows:Let $c_j$(where j = 1, 2, …, c) be a set of categorical attributes.

$$IL_c = |\gamma| \sum_{j=1}^{c} \frac{H\left(\Delta(\cup c_j)\right)}{H(\zeta\, c_j)} \quad (2)$$

Systematic clustering approach [11] from this approach they got motivation and research on information loss problem of k-anonymization introduced these approaches Approach-1: Unequal Group of QI and SA; and Approach-2: Equal Group of QI and SA and they had two objectives first was decrease information loss and second was to reserve the privacy by presenting the least number of attributes in the anonymized dataset.

- In paper [1] two approaches give minimum information loss but after that some Information loss is still there so our main goal is to decrease information loss as much as possible by 2- Level Anonymization approach in this approach anonymize only that cluster which have more information loss.
- This approach is improved version of novel approaches for privacy preserving data mining in k-anonymity model which is invented by authors Pawan Baladhare and Devesh Jinwala[1]
- Let's see how this improved version of this approach will give minimum information loss by same hospital dataset example.

## 3 COMPARISON OF DIFFERENT APPROACHES

I. Traditional K-Anonymized Approach [2]
II. Novel Approach Equal group of QI and SA [1]
III. 2-Level K-Anonymization Approach
IV.

### I. TRADITIONAL K-ANONYMIZED APPROACH VERSION OF ABOVE HOSPITAL'S DATASET SHOWN IN TABLE II

***TABLE II***
*SUBDATABASE FROM HOSPITAL'S ANONYMIZED DATASET*

| *ID* | *AGE* | *GENDER* | *ZIP CODE* | *DISEASE* |
|---|---|---|---|---|
| *3* | *[15-30]* | *P* | *132**** | *CANCER* |
| *7* | *[15-30]* | *P* | *132**** | *DIABETES* |
| *6* | *[15-30]* | *P* | *132**** | *HIGH BP* |
| *4* | *[15-30]* | *P* | *132**** | *DIABETES* |
| *2* | *[15-30]* | *P* | *132**** | *HIV* |
| *1* | *[15-30]* | *P* | *132**** | *FLU* |
| *9* | *[15-30]* | *P* | *132**** | *HIV* |
| *5* | *[15-30]* | *P* | *132**** | *CANCER* |
| *8* | *[41-50]* | *P* | *13***** | *CANCER* |
| *16* | *[41-50]* | *P* | *13***** | *HIGH BP* |
| *12* | *[41-50]* | *P* | *13***** | *HIV* |
| *10* | *[41-50]* | *P* | *13***** | *CANCER* |
| *15* | *[41-50]* | *P* | *13***** | *CANCER* |
| *11* | *[41-50]* | *P* | *13***** | *FLU* |
| *13* | *[41-50]* | *P* | *13***** | *CANCER* |
| *14* | *[41-50]* | *P* | *13***** | *DIABETES* |

Information Loss Calculation of Traditional K-Anonymization Approach for table II.

$$IL = IL_n + IL_c = 8 \times \left[\left(\frac{30-15}{50-15}\right) + \left(\frac{50-41}{50-15}\right) + (1+1) + (3+4)\right] \approx 77.50$$

II. Novel Approach Equal group of QI and SA

***TABLE III***
*K-ANONYMIZATION APPROACH EQUAL GROUP OF QA,SA(AGE,DISEASE)*

| *ID* | *AGE* | *DISEASE* |
|---|---|---|
| *3* | *[15-30]* | *CANCER* |
| *7* | *[15-30]* | *DIABETES* |
| *6* | *[15-30]* | *HIGH BP* |
| *4* | *[15-30]* | *DIABETES* |
| *2* | *[15-30]* | *HIV* |
| *1* | *[15-30]* | *FLU* |
| *9* | *[15-30]* | *HIV* |
| *5* | *[15-30]* | *CANCER* |
| *8* | *[41-50]* | *CANCER* |
| *16* | *[41-50]* | *CANCER* |
| *12* | *[41-50]* | *HIV* |
| *10* | *[41-50]* | *CANCER* |
| *15* | *[41-50]* | *CANCER* |
| *11* | *[41-50]* | *FLU* |
| *13* | *[41-50]* | *CANCER* |
| *14* | *[41-50]* | *HIGH BP* |

***TABLE IV***
*K-ANONYMIZATION APPROACH EQUAL GROUP OF QA,SA(GENDER,DISEASE)*

| *ID* | *ZIP CODE* | *DISEASE* |
|---|---|---|
| *1* | *13201** | *FLU* |
| *6* | *13201** | *HIGH BP* |
| *11* | *13201** | *FLU* |
| *3* | *13201** | *CANCER* |
| *4* | *132012* | *DIABETES* |
| *8* | *132012* | *CANCER* |
| *9* | *132012* | *HIV* |
| *14* | *132012* | *DIABETES* |
| *7* | *132050* | *DIABETES* |
| *2* | *132150* | *HIV* |
| *5* | *132150* | *CANCER* |
| *10* | *132150* | *CANCER* |
| *15* | *13***** | *CANCER* |
| *16* | *13***** | *HIGH BP* |
| *13* | *13***** | *CANCER* |
| *12* | *13***** | *HIV* |

Information Loss Calculation of Novel Approach for table III,IV,V

$IL = IL_n + IL_{c1} + IL_{c2} = 8 \times \left[\left(\frac{30-15}{50-15}\right) + \left(\frac{50-41}{50-15}\right) + (1+0) + (1+4)\right] \approx 53.50$

***TABLE V***

*K-ANONYMIZATION APPROACH EQUAL GROUP OF QA,SA(ZIP CODE,DISEASE)*

| ID | GENDER | DISEASE |
|---|---|---|
| 2 | F | HIV |
| 3 | F | CANCER |
| 5 | F | CANCER |
| 8 | F | CANCER |
| 11 | P | FLU |
| 14 | P | DIABETES |
| 16 | P | HIGH BP |
| 1 | P | FLU |
| 4 | M | DIABETES |
| 6 | M | HIGH BP |
| 7 | M | DIABETES |
| 9 | M | HIV |
| 10 | M | CANCER |
| 12 | M | HIV |
| 13 | M | CANCER |
| 15 | M | CANCER |

### III. 2-level K-Anonymization Approach

***TABLE VI***

*2-LEVEL K-ANONYMIZATION APPROACH EQUAL GROUP OF QA,SA(AGE,DISEASE)*

| ID | AGE | DISEASE |
|---|---|---|
| 3 | [15-20] | CANCER |
| 7 | [15-20] | DIABETES |
| 6 | [15-20] | HIGH BP |
| 4 | [15-20] | DIABETES |
| 2 | [21-30] | HIV |
| 1 | [21-30] | FLU |
| 9 | [21-30] | HIV |
| 5 | [21-30] | CANCER |
| 8 | [41-50] | CANCER |
| 16 | [41-50] | CANCER |
| 12 | [41-50] | HIV |
| 10 | [41-50] | CANCER |
| 15 | [41-50] | CANCER |
| 11 | [41-50] | FLU |
| 13 | [41-50] | CANCER |
| 14 | [41-50] | HIGH BP |

***TABLE VII***

*2-LEVEL K-ANONYMIZATION APPROACH EQUAL GROUP OF QA,SA(GENDER,DISEASE)*

| ID | GENDER | DISEASE |
|---|---|---|
| 2 | P | HIV |
| 3 | P | CANCER |
| 5 | P | CANCER |
| 8 | P | CANCER |
| 11 | P | FLU |
| 14 | P | DIABETES |
| 16 | P | HIGH BP |
| 1 | P | FLU |
| 4 | M | DIABETES |
| 6 | M | HIGH BP |
| 7 | M | DIABETES |
| 9 | M | HIV |
| 10 | M | CANCER |
| 12 | M | HIV |
| 13 | M | CANCER |
| 15 | M | CANCER |

***TABLE VIII***

*2-LEVEL K-ANONYMIZATION APPROACH EQUAL GROUP OF QA,SA(ZIP CODE,DISEASE)*

| ID | ZIP CODE | DISEASE |
|---|---|---|
| 1 | 13201* | FLU |
| 6 | 13201* | HIGH BP |
| 11 | 13201* | FLU |
| 3 | 13201* | CANCER |
| 4 | 13201* | DIABETES |
| 8 | 13201* | CANCER |
| 9 | 13201* | HIV |
| 14 | 13201* | DIABETES |
| 7 | 13**** | DIABETES |
| 2 | 13**** | HIV |
| 5 | 13**** | CANCER |
| 10 | 13**** | CANCER |
| 15 | 13**** | CANCER |
| 16 | 13**** | HIGH BP |
| 13 | 13**** | CANCER |
| 12 | 13**** | HIV |

Information Loss Calculation of 2-Level K-Anonymization for table VI,VII,VIII

$IL = IL_n + IL_{c1} + IL_{c2}$

$= \left[\left[4 \times \left(\frac{20-15}{50-15}\right) + 4 \times \left(\frac{30-21}{50-15}\right) + 8 \times \left(\frac{50-41}{50-15}\right)\right] + [4 \times (0+1) + 8 \times (0)] + [4 \times (1+0+0+4)]\right] \approx 27.65$

***TABLE IX***

*COMPARISION OF ALL THREE APPROACHES*

| No | Approach | Information Loss |
|---|---|---|
| 1 | Traditional K-Anonymization Approach [2] | 77.50 |
| 2 | Novel Approach K Anonymization of equal combination of QA, SA [1] | 53.50 |
| 3 | 2-level K-Anonymization Approach of equal combination of QA, SA | 27.65 |

2-level K-Anonymization Approach has lowest Information Loss that is approx. 27.65

## 4 2 – LEVEL K-ANONYMIZATION APPROACH

In table II anonymized table gives high rated information loss but in novel approach of equal group of QA and SA gives less information loss as we can see in table IV in attribute gender only 8 records get generalization instead of all and in table V in attribute zip code values got lesser suppression as compared to traditional approach. Now by comparing this approach to 2-level anonymization approach in table VII in attribute gender only 4 records generalizations, it saves 4 more to get generalized and table VIII in attribute zip code values less suppressed than both traditional and in novel approach of equal group of QA and SA approaches.

**Algorithm:**
Input: Dataset D which has r tuples
Output: γ={$\sigma 1, \sigma 2, \sigma 3, \ldots, \sigma p$} be a subdividing of r
// D is main database
// r is the amount of tuples in the dataset
// γ is a subdividing of r tuples
// $\sigma$is a clusters
// k is no of tuples for anonymization

**Begin**

1. Recognize the attributes such as identifier, quasi-identifier (QI) (numeric and non-numerical means categorical) and sensitive attribute (SA)
2. Eliminate the identifier attribute and exchange it with ID
3. Sort all tuples by their quasi-identifiers attributes.
4. Recognize the amount of equivalence classes and groups
5. Create an Equal/Unequal grouping of QI and SA to generate the sub-database
6. Make a divider of all tuples into k tuples groups
7. Select a tuple ri arbitrarily from the initial block of k tuples
8. Likewise select next tuples rj from the other block of k tuples
9. Do generalization and Suppression.
10. Do further clustering only that cluster which have more information loss
    K =k/2
11. Do generalization and Suppression.
12. Compute info loss

$$IL(\gamma) = \sum_{i=1}^{p} IL(\sigma i)$$

13. Transfer the tuples in a group with lowest information loss
14. Search additional element in a group those who surpass the k size
15. Add further element in a group whose info loss is lowermost

**End**

This algorithm steps taken from novel approach of equal/unequal group of QA and SA and step 10 and 11 are added by us to minimize information loss purpose Conferring to the algorithm we initial identify and categorize the attributes such as identifier, quasi-identifier and sensitive attributes in a dataset (step 1). Next, we eliminate the identifier attribute from the dataset and sort all tuples using the quasi-identifiers (steps 2 and 3). Then, we find out the number of groups and clusters such that $\sigma=\frac{r}{k}$, where r is the number of tuples in a database and k is the anonymization factor (step 4). Later recognizing the groups and clusters in an original dataset, we generate a sub-dataset using a grouping of QI and SA (step 5). In Approach, we create an equal/unequal grouping of QI, SA. From each generated sub-dataset, we make a partition of all tuples into k groups (step 6). Next, we used Systematic clustering algorithm in demand to generate the clusters [11]. Conferring to the Systematic clustering algorithm [11], we arbitrarily select a tuple from the first cluster for the formation of the first cluster (step 7). Likewise, we make the remaining cluster by arbitrarily selecting the tuples from the outstanding groups (step 8). Generalization and suppression applied (step 9) after that do further clustering of only that clusters which have more information loss (step 10) and again generalization and suppression. Next, we compute the information loss of each cluster (step 11). Now, we select other tuples from the initial group and add tuples in a cluster whose info loss is the lowermost (step 12). In the similar way, we select and add other tuples in a cluster whose information loss is the lowermost. Throughout the clustering process if nearly cluster has exceeded to the k size, the extra record should be added in a cluster whose info loss is the lowest (step 13 and step 14).

## 5 EXPERIMENTAL SETUP

The paper uses the ADULT dataset from the UCI Machine Learning Repository [18] for testing. The ADULT dataset holds 32561 records and 15 attributes. Out of them, we recollect only attributes Age, Gender, fnlwt, Occupation, Marital-status, Race. The attributes Age and fnlwt are numeric attributes, and Race, Gender, Marital-status and Occupation are the categorical attributes. The attribute Occupation is reserved as a sensitive attribute in the dataset. The research will be executed in Java with JDK 1.6 in a system constructed with Intel core i3 processor, 4 GB RAM and 500GB hard disk

## 6 METHODOLOGY AND EVALUATION

We run our proposed approach on various k values such as 10, 20, 30,.. up to 100. The total information loss and the execution time were calculated during each run of the experiment in fig.3. Information loss with respect to various values of k it shows when k size increase information loss respectively increases and if we compare proposed approach with previous one approach it gives lesser information loss.


***Fig.3**. Information loss VS K*

| *K* | *Traditional approach Information Loss* | *Proposed approach Information Loss* |
|---|---|---|
| *10* | *118.0625* | *37.10409* |
| *20* | *258.3561* | *84.45095* |
| *30* | *358.0526* | *113.753* |
| *40* | *518.9251* | *169.8436* |
| *50* | *648.6753* | *214.7358* |
| *60* | *777.2961* | *261.4113* |
| *70* | *909.3743* | *300.6725* |
| *80* | *1039.355* | *339.7891* |
| *90* | *1166.45* | *395.424* |
| *100* | *1297.763* | *438.5466* |

In Fig. 4. Time for execution with respect to various values of k it shows when k size increase time for execution respectively decreases Comparison between proposed approach and previous approach. Proposed approach takes bit long time because of step 10, 11 so, if we are not concern about time

then 2 level K anonymization approach is best because it gives lesser information loss.

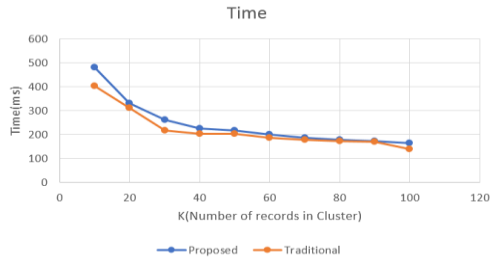


***Fig. 4*** *Execution Time*

| *K* | *Traditional approach Time(ms)* | *Proposed approach Time(ms)* |
|---|---|---|
| *10* | *403* | *481* |
| *20* | *313* | *332* |
| *30* | *219* | *262* |
| *40* | *203* | *226* |
| *50* | *203* | *219* |
| *60* | *188* | *201* |
| *70* | *180* | *188* |
| *80* | *172* | *180* |
| *90* | *170* | *172* |
| *100* | *141* | *165* |

## 7 CONCLUSION

Most of the researchers focuses only on the privacy preserving without focuses on the information loss. In this paper we focuses on both aspect privacy preserving as well as information loss because information loss is the main important issue for this anonymization so, for that 2-level anonymization approach introduced for in case of k=8 and 4. k=4 has less information loss as compared to k=8 and higher data utility but by comparing privacy parameter of both this case, k=8 have higher privacy than k=4 by this 2-level anonymization is approach which gives benefits of both first privacy parameter from k=8 and second less information loss from k=4.

## 8 REFERENCES

[1] Pawan Baladhare and Devesh Jinwala,"Novel approaches for privacy preserving data mining in k-anonymity model" Journal of information science and engineering 32,63-78(2016)

[2] Samarati, Pierangela; Sweeney, Latanya (1998). "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression" . Harvard Data Privacy Lab. Retrieved April 12, 2017

[3] Y. Lindell and B. Pinkas, "Privacy preserving data mining," Journal of Cryptology, Vol. 15, 2002, pp. 177-206.

[4] M. Upmanyu, A. M. Namboodiri, K. Srinathan, and C. V. Jawahar, "Efficient privacy preserving k-means clustering," Intelligent and Security Informatics, LNCS, Vol. 6122, 2010, pp. 154-166.

[5] G. Jagannathan, K. Pillaipakkamnatt, R. N. Wright, and D. Umano, "Communication-efficient privacy preserving clustering," Transactions on Data Privacy, Vol. 3, 2010, pp. 1-25.

[6] L. Sweeney, "k-Anonymity: a model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, 2002, pp. 557-570.

[7] J. W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques," in Proceedings of International Conference on DatabaseSystems for Advanced Applications, 2007, pp. 188-200.

[8] G. Loukides and J. Shao, "Capturing data usefulness and privacy protection in k anonymization," in Proceedings of ACM Symposium on Applied Computing, 2007, pp. 370-374.

[9] C.-C. Chiu and C.-Y. Tsai, "A k-anonymity clustering method for effective data privacy preservation," in Proceeding of the 3rd International Conference on AdvancedData Mining and Application, Vol. 4632, 2007, pp. 89-99.

[10] J.-L. Lin and M.-C. Wei, "An Efficient clustering method for k-anonymization," in Proceeding of International Workshop on Privacy and Anonymity in InformationSociety, 2008, pp. 46-50.

[11] M. E. Kabir, H. Wang and E. Bertino, "Efficient systematic clustering method for k-anonymization," Acta Informatica, Vol. 48, 2011, pp. 51-66.

[12] X. Xiao and Y. Tao, "Anatomy: simple and effective privacy preservation," in Proceedings of the 32nd International Conference on Very Large Data Bases, 2006, pp.139-150.

[13] Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: privacy beyond k-anonymity," in Proceedings of the 22nd International Conferenceon Data Engineering, 2006, pp. 1-12.

[14] N. Li, T. Li and S. Venkatasubramanian, "t-closeness: privacy beyond k-anonymity and l-diversity," International Conference on Data Engineering, 2007, pp. 106-115.

[15] R. C.-W. Wong, J. Li, A.W.-C. Fu, and K. Wang, "(a, k) anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in Proceedings of the12th ACM SIGKDD International Conference on Knowledge Discovery and DataMining, 2006, pp. 754-759.

[16] J. Goldberger and T. Tassa, "Efficient anonymization with enhanced utility," Transactions on Data Privacy, Vol. 3, 2010, pp. 149-175.

[17] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: efficient full domain k-anonymity," in Proceedings of the ACM SIGMOD International Conference onManagement of Data, 2005, pp. 49-60.

[18] UCI machine learning repository, http://archive.ics.uci.edu/ml/datasets.html.

[19] R. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in Proceedings of the 21st International Conference on Data Engineering, 2005, pp. 217-228.

[20] Analysis of Various Sentiment Classification Techniques

[21] Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis

[22] Similarity Measures for Collaborative Filtering to Alleviate the New User Cold Start Problem

[23] An Algorithmic Approach for Recommendation of Movie Under a New User Cold Start Approach
[24] Privacy Preserving by Anonymization Approach
[25] Vaghela, Vimalkumar B., and Bhumika M. Jadav. "Analysis of various sentiment classification techniques." International Journal of Computer Applications 140.3 (2016).