Minimal Cost K-Anonymization Techniques for Securing Sensitive Information

Teshu Knat Parkar
M.Tech Scholar, CSE, Shri Shankaracharya Group of Institution, Bhilai (C.G.) India

Yamini Chouhan
Assistant Professor, Department of Computer Science and Engineering, Shri Shankaracharya Group of Institution,Bhilai (C.G.) India.

Samta Gajbhiya
Professor, Head of Department, Computer Science and Engineering, Shri Shankaracharya Group of Institution,Bhilai(C.G.), India.

Abstract – There is increasing pressure to share social network data and even make it openly accessible. In any case, such revelations of individual medical data raise genuine security concerns. To mitigate such concerns, it is conceivable to anonymize the information before exposure. One well known anonymization approach is k-namelessness. There have been no assessments of the real re-distinguishing proof likelihood of k-anonymized informational datasets. In this paper, we shows two Anonymization systems for including vertices and edges for covering up valuable data. The calculation for including edges executes speeder when contrasted with calculation for including vertices yet concealing data capacity is more in vertices including calculation since it include vertices inferred that it additionally include edges so it turns out to be more strong.

Index Terms – Anonymity techniques, anonymity models, Social Network Data, privacy preserving algorithm.

1. INTRODUCTION

The k-anonymity model assumes that person-specific data are stored in a table (or a relation) of columns (or attributes) and rows (or records). The procedure of anonymizing such a table begins with removing all the unequivocal identifiers, for example, name and SSN, from the table. The details in the table by selecting only some of the columns, we can easily identify the complete row of the dataset. For e.g. Most people in the united State can be identified by a set of columns name such as {GENDER, DOB, and ZIP}

Subsequently, regardless of whether each quality alone isn't sufficiently particular to recognize people, a gathering of specific properties together may distinguish a specific person. The arrangement of such qualities is called semi identifier. The principle target of the k-anonymity demonstrate is in this manner to change a table with the goal that nobody can make high-likelihood relationship between records in the table and the comparing elements.

To accomplish this objective, the k-anonymity show requires that any record in a table be indistinct from in any event (k−1) different records regarding the pre-decided semi identifier. A gathering of records that are undefined to each other is frequently alluded to as an identicalness class. By upholding the k-anonymity prerequisite, it is ensured that despite the fact that an enemy knows that a k-unknown table contains the record of a specific individual and furthermore knows a portion of the semi identifier characteristic estimations of the individual, he/she can't figure out which record in the table compares to the person with a likelihood more noteworthy than 1/k. For instance, a 3-mysterious adaptation of the table in Fig. 1 is appeared in Fig. 2.

ZIP	Gender	Age	Diagnosis
47918	Male	35	Cancer
47906	Male	33	HIV+
47918	Male	36	Flu
47916	Female	39	Obesity
47907	Male	33	Cancer
47906	Female	33	Flu

Fig. 1. Patient Table

ZIP	Gender	Age	Diagnosis
4791∗	Person	[35-39]	Cancer
4790∗	Person	[30-34]	HIV+
4791∗	Person	[35-39]	Flu
4791∗	Person	[35-39]	Obesity
4790∗	Person	[30-34]	Cancer
4790∗	Person	[30-34]	Flu

Fig. 2. Anonymized Patient Table

2. K-ANONYMITY

k-Anonymity could be a formal model of protection [16]. The objective is to frame each record unclear from an illustrated variety (k) records if tries region unit made to detect the data. An arrangement of data is k-anonymized if, for any record with a given arrangement of characteristics, there square measure in any event k-1 elective records that match these traits. The properties can be any of the accompanying sorts.

Example:

In the event that the previously mentioned table is to be anonymized with Anonymization Level (AL) set to 2 and the arrangement of Quasi identifiers as QI = {AGE, SEX, ZIP, PHONE}.Sensitive trait = {SALARY}. The quasi identifiers and touchy qualities are distinguished by the association as indicated by their rules and regulation.

TABLE I: Table to be Anonymized

ID	Age	Sex	Zip	Phone	Salary (in Rs.)
1	24	M	641015	9994258665	78000
2	23	F	641254	9994158624	45000
3	45	M	610002	8975864121	85000
4	34	M	623410	7456812312	20000

TABLE II: Anonymized Table

ID	Age	Sex	Zip	Phone	Salary (in Rs.)
*	20-50	ANY	641***	999*******	78000
*	20-50	ANY	641***	999*******	45000
*	20-50	ANY	612***	897*******	85000
*	20-50	ANY	623***	745*******	20000

A. Generalization

Generalization is the way toward changing over an incentive into a less particular general term. For ex, "Male" and "Female" can be generalized to "Any". At the accompanying levels generalization procedures can be connected.

B. Suppression

Suppression comprises in averting delicate information by evacuating it. Suppression can be connected at the level of single cell, whole tuple, or whole segment, permits diminishing the measure of speculation to be forced to accomplish k-anonymity.

3. LITERATURE SURVEY

Xuyun Zhang et al. [16] proposes giving security and assurance over the moderate informational collections progress toward becoming question issue since enemies may hold smaller scale information by distinguishing numerous information records. Encryption of all datasets as a rule society arrange called cloud take in past frameworks may to a great degree monotonous and excessive.

Mohammad Reza Zare Mirakabad et al. [17] focuses giving insurance over the data creation. Under security data use and abhorrence of disclosure of individual identity is more basic. One of the data anonymization strategies called K-mystery keeps the revelation of individual character anyway it is generally fail to achieve.

Min Wu et al. [18] proposes sparing security is most essential however a comparable time it is bother in landing of little scale data release. In the point of view of quality divulgence K-anonymity isn't well. So we propose new framework called an ordinal partition based affectability mind full varying characteristics metric model.

Yunli Wang et al. [19] proposes k-anonymity fails to achieve characteristics disclosure however in l-grouped characteristics intends to achieve trademark presentation. Second data anonymization method center around cutting the illation from freed scaled down scale characteristics.

Jordi Soria Comas et al. [20] focuses data anonymization techniques spare assurance, k-anonymity and €-differential security are two rule insurance show. The t-closeness is the increase of k-indefinite quality, the advancement of private touchy information relies upon Bucketization calculation.

4. METHODOLOGY

In this section we present the two Anonymization techniques

a. Anonymization using graph vertices

b. Anonymization with graph edges

C. Anonymization Using Graph Vertices

In this the vertices are anonymized. The extra vertices are intentionally added to the network or chart to conceal the level of data show in it.

D. With Edges Anonymization

In this the edges are anonymized. The extra edges are intentionally added to the network or chart to conceal the level of data show in it.

Including vertices and edges are relied on the k-anonymization algorithm.

For Edges Anonymization – Edges Anonymization calculation is utilized. The calculation successfully anonymized the information. The algorithm is displayed in fig. 2.

For Vertex Anonymization – Vertex Anonymization algorithm is used. The algorithm effectively anonymized the data. The algorithm is presented in fig. 3.

Algorithm: Edges Anonymization

Input: An initial multi sensitive graph G(V, E)

Output: Graph G'(V', E') – with added edges

1. Get degrees in descending order
2. Anonymize the degree
3. Using anonymize vector, add additional degree
4. Create subgraph for the degree vector
5. Return Anonymized graph

Fig. 2. Shows the algorithm of E-Anonymization

Algorithm: Vertices Anonymization

Input: An initial multi sensitive graph G(V, E)

Output: Graph G'(V', E') – with added vertices

1. fetch orbits from the graph using stab graph Algorithm
2. Iterate through the orbits of the graph
   a. Introduce new vertex and add to the graph and include it into orbit
   b. Get ID of the vertex
   c. Connect new edges according to the orbit
   d. In same orbit connect them by tag and in different connect them by the regular graph connection
3. Return Anonymized graph

Fig. 3. Shows the algorithm of V-Anonymization

For adding least number of edges in the graph the following equation is used:

If

$$L_1\left(\hat{\mathbf{d}}-\mathbf{d}\right)=\sum_i\left|\hat{\mathbf{d}}(i)-\mathbf{d}(i)\right|,$$

Then the minization of egdes can be converted into the problem of minimization of L_1 distance of seuqences of degree of G and G'. Based on quation:

$$G_A(\hat{G},G)=\left|\hat{E}\right|-|E| \;=\; \frac{1}{2}L_1\left(\hat{\mathbf{d}}-\mathbf{d}\right).$$

Where G is the Graph with E edges and V vertices.

d is the distance minimum from the source. And G' is the sugraph with E' and V' with edges and vertices respectively.

5. RESULTS

The analysis are conducted utilizing Eclipse framework on java environment. We have exhibited two calculation for anonymization. First of all by including edges and besides by including vertices. As including vertices is extremely mind boggling process since we can't just include vertices into the graph, we require additionally to include edges as a matter of course. Consequently time taken by the including vertices is long when contrasted with including edges. Fig. 4. Demonstrates the points of interest of Facebook network dataset.

Degree	Vertices
1	75
2	98
3	93
4	99
5	93
6	98
7	98
8	111

Total Vertices	4039
Vertices added (%)	0 (0.00%)
Total Edges	88234
Edges added (%)	0 (0.00%)
Duration	0.00sec

Fig. 4. Snapshot of social circle dataset

Table: III. 5-Anonlymized – Adding Edges Output

Total Vertices	**4039**
Vertices Added	0
Total Edges	95420
Edges Added	7186 (8.14%)
Time Taken	3.59 sec

Table: IV. 5-Anonlymized – Adding Vertices Output

Total Vertices	**4386**
Vertices Added	347 (8.59%)
Total Edges	181487
Edges Added	93253 (105.69%)
Time Taken	16.65 sec

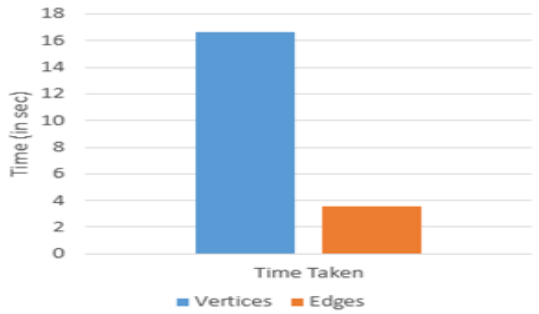

Fig. 5. Shows the time taken for execution of two algorithms

6. CONCLUSION

In this paper, we proposed a productive k-anonymization calculation by changing the k-anonymity issue to the k-part hiding issue. We likewise proposed two imperative components of hiding information, that is, edge based and other is vertices based. We performed experiment, and the result shows that, by using edges none vertices are includes that means it takes less time. But when using vertices anonymization, large number of edges are added to the graph and hence takes more time. If there are large number of information to hide, the edges anonymization is efficient, since it takes less time to hide information.

REFERENCES

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

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

[3] 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.

[4] Xuyun Zhang, Chang Liu, Surya Nepal, Chi Yang, Wanchun Dou, Jinjun Chen" Combining Top-Down and Bottom-Up: Scalable Sub-Tree anonymization over Big data using MapReduce on Cloud".

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

[6] M. Terrovitis, N. Mamoulis, and P. Kalnis. "Privacy-preserving anonymization of set-valued data." PVLDB, 1(1):115–125, 2008.

[7] Md Nurul Huda, Shigeki Yamada, and Noboru Sonehara, "On Enhancing Utility in k-Anonymization", International Journal of Computer Theory and Engineering, Vol. 4, No. 4, August 2012.

[8] Pawan R. Bhaladhare and Devesh C. Jinwala, "Novel Approaches for Privacy Preserving Data Mining in k-Anonymity Model" , JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 32, 63-78 (2016).

[9] Mohammed, N. and Fung, B. C. M, "Centralized and distributed anonymization for high-dimensional healthcare data", ACM Trans. Knowl. Discov. Data. 4, 4, Article 18 (October 2010), 33 pages.

[10] S. E. Fienberg, A. Slavkovic and C. Uhler," Privacy Preserving GWAS Data Sharing", 2011 11th IEEE International Conference on Data Mining Workshops.

[11] A. G. Divanis and G. Loukides," PCTA: Privacy-constrained Clustering-based Transaction

[12] Data Anonymization", ACM 2011.

[13] S. Kisilevich, L. Rokach, Y. Elovici, and B. Shapira. "Efficient multidimensional suppression for k-anonymity." TKDE, 22:334–347, 2010.

[14] G. Loukides, A. Gkoulalas-Divanis, and B. Malin. Anonymization of electronic medical records for validating genome-wide association studies. PNAS, 17:7898–7903, 2010.

[15] J. Cao, P. Karras, C. Ra¨ıssi, and K. Tan. rho-uncertainty: Inference-proof transaction anonymization. PVLDB, 3(1):1033–1044, 2010.

[16] Loukides, G., Shao, J.: Capturing data usefulness and privacy protection in k-anonymisation. In: Proceedings of the 2007 ACM Symposium on Applied Computing (2007)

[17] X. Zhang, C. Liu, S. Nepal, S. Pandey and J. Chen, "A Privacy Leakage Upper Bound Constraint-Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 6, pp. 1192-1202, 2013.

[18] Mohammad Reza Zare Mirakabad School of Computer Sciences, USM, Malaysia Intern at School of Computing, NUS, Singapore reza@cs.usm.my, reza.z@comp.nus.edu.sg "Diversity versus Anonymity for Privacy Preservation".

[19] Min Wu, Xiaojun Ye Institute of Information System and Engineering School of Software, Tsinghua University, Beijing, 100084, China "Towards the Diversity of Sensitive Attributes in k-Anonymity".

[20] Yunli Wang, Yan Cui, Liqiang Geng and Hongyu Liu, "A new perspective of privacy protection: Unique distinct l-SR diversity," 2010 Eighth International Conference on Privacy, Security and Trust, Ottawa, ON, 2010.