ISSN 2320 - 2602
Volume 3, No.1, January 2014
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2014/ijacst05312014.pdf

# Analysis of *k*-Anonymity for Homogeneity Attack

**Debasis Mohapatra[1], Dr. Manas Ranjan Patra[2]**
[1]PMEC, India, devdisha@gmail.com
[2]Berhampur University, India, mrpatra12@gmail.com

## ABSTRACT

The size of data is increasing exponentially day by day. It is as much important to provide privacy as to provide utility. Objective of privacy preserving data mining is to serve knowledge extraction with out leakage of private sensitive information, k-anonymity approach is a group based anonymization approach devised to achieve this objective. In this paper (α,k)-anonymization is discussed, that is a performance analysis metric, that finds the degree of association of sensitive value with an equivalence class. An algorithm is designed to return α value for each sensitive attribute. This paper represents a complete bipartite graph representation of (α,k)-anonymization, that can also generate accurate α values by implementing the algorithm on the graph. Performance analysis of k-anonymization is discussed with respect to the effects of homogeneity attack based on different parameters.

**Keywords:** k-anonymity, privacy preserving data mining, (α,k)-anonymization, bipartite graph .

.

## 1. INTRODUCTION

Now a day we are dealing with a massive amount of data that contain person-specific information. From the person-specific data it is easy to extract the sensitive information about a person by the process of linking. Today data privacy is having equal importance as data utility. *k*-anonymity[1,8,9,11] is a formal approach to prevent the database release from linking and could able to provide privacy preserving data mining. If two firms are interested to apply the data mining on the union of their data set then their intension is to extract good data mining result without disclosing their private data, *k*-anonymity can be used as a tool to achieve the above requirement. The prime concern in privacy preserving data mining is not to reveal sensitive information. There are various approaches like secure multiparty computation[3,4,6] and randomization are present to provide privacy preserving data mining. *k*-anonymity is a group based anonymization approach to provide the same. In this paper we show the basic *k*-anonymity method and performance analysis of *k*-anonymity with respect to different changes in parameters. In this paper we explain (α,*k*)-anonymity that shows the degree of association of a sensitive value with all equivalence classes, α is chosen as the maximum value among these degree of association, it is used to measure the performance of *k*-anonymity, also useful in designing enhanced *k*-anonymity. We explain the algorithm for (α, *k*)- anonymity . Also we elucidate the bipartite graph representation of (α, *k*)- anonymity . There are two attacks on *k*-anonymity: homogeneity attack and background knowledge attack[7]. *k*-anonymization does not provide protection against background attack[7]. In this paper we explain the effect of homogeneity attack when there is a change in *k* value but database size is fixed and change in data base size with a fixed *k* value.

## 2. RELATED WORK

L. Sweeney [11] has introduced a formal protection model called *k*-anonymity and a set of accompanying policies for deployment, also examines re-identification attacks. L. Sweeney [10] has proposed generalization and suppression methods to achieve *k*-anonymity. J.D. Ferrer et al. [5] have proposed an approach to use categorical microaggregation as an alternative to generalization and suppression for nominal and ordinal *k*-anonymization. R.C. Wong et al.[12] have dicussed (α,k)-anonymity model and also proved that the optimal(α,k)-anonymity problem is NP-hard. A. Machanavajjhala et al. [7] have discussed the *k*-anonymity with homogeneity attack and background attack and proposed *l*-diversity model.

## 3. *k*-ANONYMITY

In *k*-anonymity approach the records or tuples are grouped into n/k equivalence classes where each equivalence class is having *k* tuples and the database is having n tuples. The k records of an equivalence class are indistinguishable according to their quasi-identifiers value. Quasi-identifiers [1,2,4,6] are the attributes that are linked with other external database to uniquely identify the entities of the table. The release table T after *k*-anonymization is converted into anonymized table $T^*$. There are different classification of attributes are present; an attribute is called as categorical if it takes nonnumeric values otherwise called as numerical, an attribute is called as sensitive if it takes sensitive information about an entity like DISEASE.

In Table 1(Hospital Data) the non sensitive attribute set NS={PID,STATE,AGE} and sensitive attribute set S={DISEASE}. Table 2 is the 4-anomymous table of Table 1, which contains 3 equivalence classes each having 4 indistinguishable tuples with respect to quasi-identifiers {PID,STATE,AGE}. Suppression[5,11] is used in STATE

attribute whereas in PID and AGE generalization[5,11] is used to provide anonymization. The main objective in this method is to provide the analysis of the sensitive attribute without disclosing the sensitive information of a particular entity. Table 1 is converted into Table 2 to preserve the privacy of sensitive information i.e. the disease of a person should not identified from the table. PID, STATE, DISEASE are the categorical attributes where as AGE is a numerical attribute.

**Table 1:** Hospital Data

| NONSENSITIVE | | | SENSITIVE |
|---|---|---|---|
| **PID** | **STATE** | **AGE** | **DISEASE** |
| 121045 | Odisha | 27 | Brain Cancer |
| 121077 | Bihar | 28 | Malaria |
| 121088 | UP | 29 | Heart Disease |
| 121067 | MP | 22 | Malaria |
| 134222 | Odisha | 55 | Heart Disease |
| 134567 | MP | 54 | Malaria |
| 134889 | UP | 50 | Heart Disease |
| 134778 | Bihar | 47 | Malaria |
| 143367 | Odisha | 35 | Heart Disease |
| 148000 | MP | 37 | Brain Cancer |
| 145690 | UP | 33 | Malaria |
| 148056 | Bihar | 34 | Brain Cancer |

**Table 2 :** 4-anonymous Hospital Data

| NONSENSITIVE | | | SENSITIVE |
|---|---|---|---|
| **PID** | **STATE** | **AGE** | **DISEASE** |
| 1210** | * | <30 | Brain Cancer |
| 1210** | * | <30 | Malaria |
| 1210** | * | <30 | Heart Disease |
| 1210** | * | <30 | Malaria |
| 134*** | * | >40 | Heart Disease |
| 134*** | * | >40 | Malaria |
| 134*** | * | >40 | Heart Disease |
| 134*** | * | >40 | Malaria |
| 14**** | * | 3* | Heart Disease |
| 14**** | * | 3* | Brain Cancer |
| 14**** | * | 3* | Malaria |
| 14**** | * | 3* | Brain Cancer |

Remark :- *Homogeneity of sensitivity attribute in the equivalence class leads to the leakage in privacy.*

Suppose 'n' tuples are present in databases that obey *k*-anonymity. There are 'm' numbers of sensitive attributes present with $K_1, K_2, \ldots\ldots K_m$ distinct values. The number of equivalence classes present that can have the possibility of all sensitive attributes value same in the particular equivalence class is say 'SE'. This value of 'SE' determines the possibility of homogeneity attack.

$SE=(K_1K_2K_3\ldots\ldots K_m)/(K_1^kK_2^kK_3^k \ldots\ldots\ldots K_m^k)*(n/k)$……(1)

Proof:- let us consider, m sensitive attributes are in the database, in each equivalence class k tuples are there, the event is homogeneity attack (HA).

$P(HA)=(K_1K_2K_3\ldots\ldots\ldots K_m)/(K_1^k K_2^k K_3^k \ldots\ldots\ldots\ldots K_m^k)$. So number of groups having same sensitive value is $P(HA)*(n/k)$.

In Table 1 n=12, $k$=4, m=1, $K_1$=3 so S=1/9=0.11111. In this case the possibility of homogeneity attack[7] is 0.11111, means 0.11111 equivalence classes having all sensitive value equal inside the equivalence class . So it is advised that along with the *k*-anonymization, sanitized table should ensure diversification of sensitive values for better performance.

## 4. (α,*k*) ANONYMITY

The database 'DB' obeys both k-anonymization and α-association[12] with respect to quasi-identifier and sensitive value then it is said to obey (α,K) anonymity**.** Let in a database 'DB', ' QI' represents the quasi-identifiers set, ' S' denotes the sensitive attribute set. Database 'DB' is α associated with respect to attribute set 'QI' and sensitive value 's' if $|(E,s)|/|E|\leq\alpha$, for all equivalence class E. It is one of the metric to analyze the performance of *k*-anonymity and a way to design enhanced *k*-anonymity. Table 3 explains the process of evaluation of different α values with respect to sensitive values S={Brain Cancer, Malaria, Heart Disease } and quasi-identifier QI={PID, STATE, AGE}.

### 4.1 Algorithm (α,*k*) Anonymity

Input:- $S=\{S_1,S_2,\ldots\ldots,S_v\}$ is the set of sensitive values, T is the database table, n=|T|.

Output:- $\alpha=\{\alpha_1, \alpha_2,\ldots\ldots\ldots \alpha_v\}$ with respect to $\{S_1,S_2,\ldots\ldots,S_v\}$

Assumption:- All the tuples of an equivalence class are placed adjacent to each other and n%k=0.

```
1. m = n/k
2. for  i = 1 to v
{
α [i] = 0
for  t = 0  to m-1
   {
     count=0
     for  j = tk+1 to (t+1)k
        {
          if(S[i] == T.S[j])
                count = count+1
        }
      a= count/k
   if (a > α [i])
     α [i] = a
   }
}
3. return  α
```

This algorithm finds the α [i] degree of association of a sensitive value $S_i$ for given *k*-anonymity.

### 4.2 Bipartite Graph Representation of (α,*k*) Anonymity

(α,*k*) anonymity can be represented by a complete bipartite graph G(S,E,W) where S ={$S_1$,$S_2$,……,$S_m$} is the set of 'm' vertices represent 'm' sensitive values, E={$E_1$,$E_2$,……..,$E_n$} is the set of 'n' vertices represent 'n' equivalence classes, W={$W_1$,$W_2$,……..,$W_{mn}$} is the set of 'mn' weighted edges between each vertex of 'S' to each vertex of 'E'. Weight of the edge $(S_i,E_j) = |(E_j,S_i)|/|E_j|$ $1 \le i \le m$, $1 \le j \le n$ . In Table 3, S={Brain Cancer, Malaria, Heart Disease } ,E= {$E_1$,$E_2$,$E_3$}. Figure 1 represents the complete bipartite graph that shows each sensitive value is attached to all equivalence classes by an edge. In Table 4 the weight of the edges are stored in cost adjacency matrix. Here Brain Cancer=BC, Malaria=M, Heart Disease=HD. α value with respect to QI and sensitive value $S_i$ is the maximum value present in the $S_i$ row of cost adjacency matrix for example (0.5, 4) anonymity with respect to QI and Brain Cancer.

Lemma:- The database 'DB' is called (α,k) anonymity with respect to QI and $S_i$ then $\alpha=\max(S_i,E_j)$ $1 \le j \le n$, in a complete bipartite graph where $S_i \in S$, $E_j \in E$.

### 4.3 Graph based Algorithm for (α,*k*) Anonymity

Input:- Graph G(S,E)S={$S_1$,$S_2$,……,$S_v$},
E={$E_1$,$E_2$,……..,$E_m$}, Cost adjacency matrix C.
Output:- α={ $\alpha_1$, $\alpha_2$,…….. $\alpha_v$ } with respect to {$S_1$,$S_2$,…,$S_v$}

1. for i = 1 to v

```
{
  a=0
   for  j = 1 to m
     {
       if(C[i,j] > a)
        a= C[i,j]
     }
 α[i]=a
}
```

2. return α

This algorithm also finds α [i] degree of association of a sensitive value $S_i$ for given *k*-anonymity, but here the algorithm is implemented on complete bipartite graph. The complete bipartite graph provides a simple and better representation of (α,*k*) anonymity.

**Table 3 :** (α,*k*) anonymity based on Table 1

| S={Brain Cancer, Malaria, Heart Disease } ,QI={PID, STATE, AGE} | | | |
|---|---|---|---|
| (E, Brain cancer) | | | |
| ($E_1$, Brain Cancer) | ($E_2$, Brain Cancer) | ($E_3$, Brain Cancer) | (E, Brain Cancer)/\|E\| ≤ 0.5 for all E (0.5,4) anonymity w.r.t QI and Brain cancer |
| \|($E_1$, Brain Cancer)\|/\|E\| = 1/4=0.25 | \|($E_2$, Brain Cancer)\|/\|E\| = 0/4=0 | \|($E_3$, Brain Cancer)\|/\|E\| = 2/4=0.5 | |
| | | | |
| (E, Malaria) | | | |
| ($E_1$, Malaria) | ($E_2$,Malaria ) | ($E_3$, Malaria) | (E, Malaria)/\|E\| ≤ 0.5 for all E (0.5,4) anonymity w.r.t QI and Malaria |
| \|($E_1$, Malaria)\|/\|E\| = 2/4=0.5 | \|($E_2$, Malaria)\|/\|E\| = 2/4=0.5 | \|($E_3$, Malaria)\|/\|E\| = 1/4=0.25 | |
| | | | |
| (E, Heart Disease) | | | |
| ($E_1$, Heart Disease) | ($E_2$, Heart Disease ) | ($E_3$, Heart Disease) | (E, Heart Disease)/\|E\| ≤ 0.5 for all E (0.5,4) anonymity w.r.t QI and Heart Disease |
| \|($E_1$, Heart Disease)\|/\|E\| = 1/4=0.25 | \|($E_2$, Heart Disease)\|/\|E\| = 2/4=0.5 | \|($E_3$, Heart Disease)\|/\|E\| = 1/4=0.25 | |

**Figure 1** :Complete Bipartite Graph

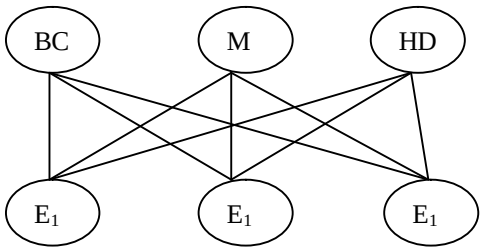
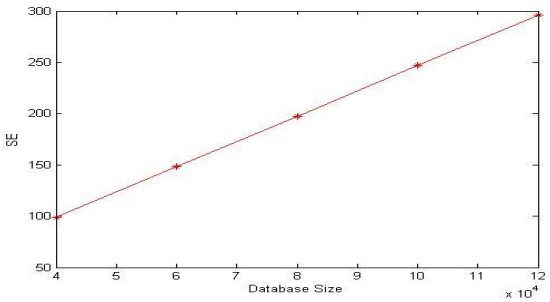



**Table 4 :** Cost Adjacency of Figure 1

| (S,E) | $E_1$ | $E_2$ | $E_3$ |
|---|---|---|---|
| BC | 0.25 | 0 | 0.5 |
| M | 0.5 | 0.5 | 0.25 |
| HD | 0.25 | 0.5 | 0.25 |

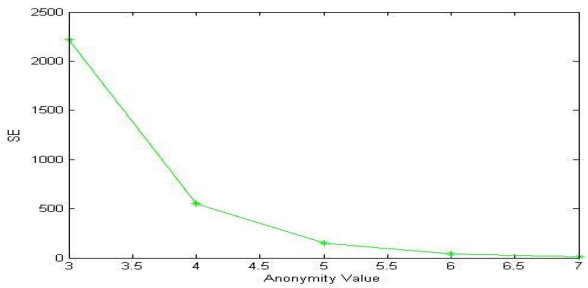
## 5. HOMOGENEITY ATTACK WITH RESPECT TO DIFFERENT PARAMETERS

Case-1:- Here we are considering the databases with different size , m (no of sensitive attribute) value is 1, 5-anonymization. Here the size of database is varying but anonymization is fixed to 5.




Observation: - It is clear that if the 'Database Size' increases with 5-anonymity fixed, the SE value increases or possibility of homogeneity attack increases.

Case-2:- Here we are considering the databases with different anonymity values represented by k, m (no of sensitive attribute)=1.Here the size of database is fixed to 60000 but anonymization is varying.




Observation: - Anonymity value is increasing with the constant database size 60000.The SE value is decreasing or possibility of homogeneity attack decreases.

In (α,k) anonymity higher the α value higher the chance of homogeneity attack. So the sanitized table should ensure least α value and the difference between the α values with respect to different sensitive values should be minimized for diminishing the possibility of homogeneity attack.

## 6. CONCLUSION

*k*-anonymity is an interesting approach of group anonymization. We have discussed the basic concepts of *k*-anonymity with homogeneity attack analysis. (α,*k*) anonymity is discussed along with its bipartite graph representation. Bipartite graph representation is able to provide an efficient simulation of (α,*k*) anonymity problem and is an effective method to analyze the α values.

## REFERENCES

1. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu, "Anonymizing tables". In *ICDT*, 2005, pages 246–258.
2. R. Agrawal, R. Srikant, " Privacy-preserving data mining ". In *Proc. of the ACM SIGMOD Conference on Management of Data*, May 2000, pages 439–450.
3. D. Agrawal, C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms". In *Proc. of ACM Symposium on Principles of Database Systems*, 2001
4. C. Clifton, M. Kantarcioglu ,J. Vaidya , Lin X. , Zhu Michael Y., " Tools for Privacy Preserving Data Mining " . *International Conference on Knowledge Discovery and Data Mining, Vol. 4, No. 2,* 2002, Pages 1-8 .
5. J.D. Ferrer, V.Torra, " Ordinal, Continuous and Heterogeneous *k*-Anonymity Through Microaggregation ", *Data Mining and Knowledge Discovery, 11,* 2005,Pages195–212.
6. Y. Lindell , B. Pinkas , " Secure Multiparty Computation for Privacy-Preserving Data Mining " , *The Journal of Privacy and Confidentiality , Number 1,* ,2009, Pages 59-98.
7. A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, "*l*-diversity: Privacy beyond *k*-anonymity", In *Proc. 22nd Intnl. Conf. Data Engg. (ICDE)*, 2006, Page 24.
8. A. Meyerson, R. Williams, "On the complexity of optimal k-anonymity". In *PODS*, 2004, pages 223–228,.
9. P. Samarati, "Protecting respondents' identities in microdata release". *IEEE Transactions on Knowledge and Data Engineering*, 13(6), 2001, pages1010-1027.
10. L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression". *Int'l Journal on Uncertainty, Fuzziness, and Knowledge-Base Systems* 10(5), 2002, pages571-588.
11. L. Sweeney , " *k*-ANONYMITY: A Model for Protecting Privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5),* 2002, Pages 557-570.
12. R. C. Wong , J. Li , A. W. Fu , K. Wang, " (α, k) Anonymity: An Enhanced kAnonymity Model forPrivacyPreserving Data Publishing", *KDD'06,* August 20–23, 2006.