Journal of Telecommunications System & Management

Rajeswari et al., J Telecommun Syst Manage 2018, 7:2
DOI: 10.4172/2167-0919.1000164

Review Article Open Access

Anonymization of Group Membership Information Using t-Closeness and Cuckoo Filters

Rajeswari N*, Sumalatha L and Suneetha Eluri

Department of Computer Science and Engineering, University College of Engineering (A) JNTU Kakinada 533003, East Godavari, Andhra Pradesh, India

Abstract

Social networks are used by more than a billion people worldwide. They are inherent part of today's Internet. A platform to build social relationships among people is social networking service (SNS). People share their activities, interests and backgrounds which incorporate new information through blogging, mobile connectivity and photo/video sharing. Information sharing leads to different kinds of attacks like friendship attack, mutual friend attack and group membership attack. Out of these attacks Group membership attack is a serious vulnerability. In this attack malware authors can easily spread malicious code and gain information about the members of the group. Hence user's privacy is a key concern. Anonymization is a technique used to control the information access gained by the membership users. In this paper, Cuckoo filters are used to perform set membership test of a group and t- closeness is used for anonymizing structural data while publishing the data in a group. And the results have shown that proposed system yields better performance in terms of risk and gain/loss.

Keywords: Anonymity; De-anonymization; Cuckoo hashing; Bloom filters; Data privacy; Micro aggregation; K-anonymity; t-Closeness

Introduction

The usage of social networks has gained more interest in recent years for publishing sensitive and non-sensitive data. Protecting the private information (sensitive) of individuals while publishing the data is necessary. Hence privacy control in the context of storage, processing and publishing of data is a major challenge and interesting research area for ensuring the utility of social networks [1,2]. Social network represented as a graph in which the nodes are users and the edges are the connections between them. In social networking different kinds of attacks are possible like friendship attack, mutual friend attack and group membership attack. In a published social network data set they are connected by an edge to re-identify related victims. In friendship attack, the adversary uses the degrees of two vertices. In mutual friend attack by using the number of mutual friends adversary can re-identify a pair of friends. Group membership attack is an active attack; it exploits the information about the members in a group, which includes number of persons bound together by the common social standards, interests, etc. An often practice to secure published data in social networks is data anonymization i.e., to remove easily identifiable attributes like names and sensitive attributes like postal codes, social security numbers (SSN), contact numbers and e-mail addresses etc., and retaining the quasi identifiers like zip codes, etc. The paper is organized as related work about t-closeness and cuckoo filters (TCF) discussed in section II, Section III explains about the architecture of TCF. The process of applying TCF to social networks is explained in Section IV. In Section V, results are evaluated and conclusion is discussed in Section VI.

Related Work

Initially set membership test [3] is performed to check whether the user is a member of a group or not. If the user is a member of group then only the user is allowed to access information about the group. Approximate set membership tests are used by many caches, databases, storage systems and routers to decide whether the given element is present in a data (usually large) set. Bloom filters are most widely-used data structure to perform this test. In this method, elements can be added to the set, but elements cannot be removed from it. For this purpose a data structure is used, a variant of cuckoo filters, which provides four major advantages when compared with the Bloom Filters [4]. It supports removing and adding values dynamically. When compared to traditional Bloom filters, it Will provide higher lookup (search) performance even though the hash table is close to full (i.e., 95% of space is used). Set membership test simply search for an item x in the hash table. If the value x is found then it returns true [5,6]. Other alternative data structures like quotient filter which utilize less space when compared with the Bloom filters also exist. However, more practical applications use cuckoo filters as they have the target false positive rate (€) generally less than 3%. And then the motivation behind data publishing is anonymization i.e., to remove the "sensitive" information before publishing. The use of social networks is preserved to a maximum extent without compromising user's privacy. In many high-profile cases, privacy is equivalent to anonymity [7]. L-diversity is a group based anonymization technique proposed in a study [8]. It preserves privacy in data sets by additionally maintaining the diversity of sensitive fields. But l-diversity is not sufficient to prevent the adversary from gaining knowledge when sensitive attributes are semantically similar. And to prevent the disclosure of attribute values when the overall distribution is skewed [9]. Hence the focus is on more sophisticated anonymization technique such as t-closeness because of its strict privacy guarantee [10]. Hence, research is still in its infancy in terms of attacks and their defense mechanisms.

System Architecture

In the proposed system, we perform set membership test to check whether a particular user is a member of a group or not. If the user exists then only the user is allowed to access information about the group.

***Corresponding author:** Rajeswari N, Department of Computer Science and Engineering, University College of Engineering (A) JNTU Kakinada 533003, East Godavari, Andhra Pradesh, India, Tel:+918374033833; E-mail: raji.nitta@gmail.com

Received July 27, 2018; **Accepted** July 30, 2018; **Published** August 06, 2018

Citation: Rajeswari N, Sumalatha L, Eluri S (2018) Anonymization of Group Membership Information Using t-Closeness and Cuckoo Filters. J Telecommun Syst Manage 7: 164. doi: 10.4172/2167-0919.1000164

This operation is carried out by Cuckoo filters (Figure 1). The concept of Cuckoo filters is based on Cuckoo hashing. In 2001, Rasmus Pagh and Flemming Friche Rodler are the first who described about it. It stores only fingerprints in cuckoo hash table which is a compact variant of Cuckoo filters. For each string insertion hash function is applied on input string to derive the bit string and is considered as fingerprint rather than using key-value pairs [5]. Data anonymization is a technique which converts the sensitive data to anonymize data, from which data cannot be recovered or disclosed to the adversary. Different data anonymization techniques are available like k-anonymity, l- diversity and differential privacy. But, attribute disclosure is not prevented by k-anonymity. It leads to different kinds of attacks such as "Background Knowledge Attack and Homogeneity Attack". To prevent attribute disclosure l-diversity is not sufficient. And it leads to "Skewness Attack and Similarity Attack". To overcome the above challenges, t-closeness principle is used for guaranteed privacy [10]. t-closeness is a group based anonymization which is further a refinement of l-diversity. Privacy of structural data set is preserved and data can be viewed in different levels. T-closeness decides the correlation between sensitive attributes and quasi identifier attributes. This correlation limits the information disclosure. In t-closeness method, it is necessary to import data and configuration and define transformation model such micro aggregation. Generalization reduces the granularity of data sets which leads to information loss [11]. Hence micro aggregation is preferred over generalization as it has several advantages like preserving the granularity of categorical and numerical data sets. Also in micro aggregation outliers are minimized and different levels of hierarchy can impose on disclosure of sensitive attributes. We defined privacy model like t-closeness which ensure better privacy than other basic group based anonymization techniques like l-diversity and k-anonymity [12]. Next define coding model such as earth mover's distance with equal ground distance or hierarchal ground distance. The results are evaluated and visualized in graphical view that fulfill a user's requirements for preserving privacy in data transformations and finally the transformed dataset is compared with the original input dataset and analyzed the risks, accuracy and export data as de-identified.

Applying TCF to Social Network Users

The following steps are used for performing set membership test. Initially hash table is maintained for registered users, say H. H[value]=true, when the user exists. H[value]=false, when the user is not existing. A. Applying Cuckoo filters for set membership test A cuckoo filter is a practical data structure based on cuckoo hashing.

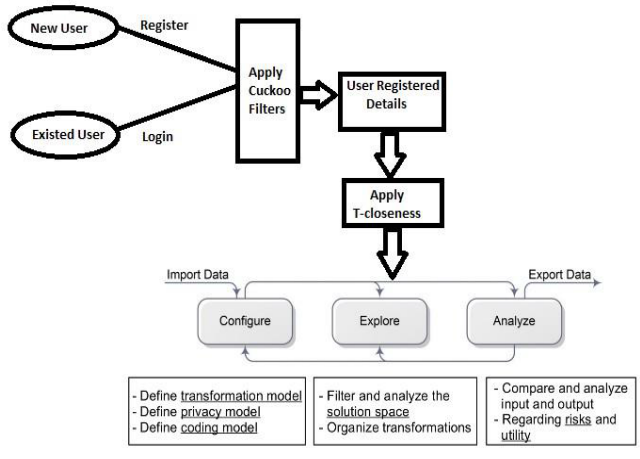

Figure 1: Architecture of TCF.

The design hash table is based on open- addressing [5]. The values are stored in an array of large size, with no linked list and pointer concept.

a) For Cuckoo insertion algorithm, two hash functions required:

Algorithm 1: Cuckoo Insert(S)

j1=hash_1(S, n);

j2=hash_2(S, n);

//S is the string of sensitive attributes and n is the//length of the string

if bucket[j1] or bucket[j2] have an empty slot then Insert S into the bucket;

Return Ok;

//need to relocate existing values;

j=randomly select j1 or j2;

for m=0 to MaxNumofKicks do pick any entry i from bucket[j];

Interchange S and the value stored in entry i;

j=hash_2(S, n);

if (any entry in bucket[j] is empty) Insert S into bucket[j];

Report Ok;

//Hash table as full;

Return Fail;

Initially, the values are stored in either of the two candidate buckets that are calculated from the hash functions. It is a memory efficient and high performance data structure used for performing set membership test.

Cuckoo filters use two hash functions, the hash table size is in powers of two and the value of item x is determined by two hash functions. For each item to be inserted, there are two candidate buckets. The search algorithm is used to check if either of bucket contains the item x. If the value is not found, then insert x by using algorithm 1. Figure 2 shows how to insert a new item (value) x into hash table of size 8 buckets. Here, x can be placed in either of the bucket 6 or 2. If either of x's two candidate buckets are empty, algorithm 1 is used to insert x to that empty bucket and declares the insertion is successful. If neither of the two candidate buckets has space, in this case, the candidate bucket is selected by its value for e.g., here bucket 6. The existing value (here 'a') is kicked out and this victim value is re-inserted to alternate location of its own.

b) For Cucko Search Algorithm, only hash table is required:

Algorithm 2: Cuckoo Search(S)

j1=hash_1(S, n);

j2=hash_2(S, n);

if bucket[j1] or bucket[j2] has S then

return

else

True;

return

False;

c) For Cuckoo Delete Algorithm, only hash table is required:

Algorithm 3: Cuckoo Delete(S)

j1=hash_1(S, n);

j2=hash_2(S, n);

if bucket[j1] or bucket[j2] has S then

Delete S from hash table;

Applying t-Closeness for Anonymization

a) **Methodology**

The structured data is given as input to t-closeness. It is used for data anonymization which ensures better privacy than l-diversity and k-anonymity [13,14]. tcloseness is the distance of the distribution of sensitive attributes to the distribution of remaining attributes of the whole class. For an equivalence class of attributes the threshlod is minimum If all the equivalence classes have t- closeness, then the structured data set is said to have t- closeness. tcloseness is the difference between global distribution of values 'G' and the distribution of sensitive attributes 'S' values within the anonimized group. The threshold 'T' gives an upper bound on it.

$$G - S < T \quad (1)$$

So, the most efficient anonymization technique t- closeness is better than many other techniques that are used for preserving privacy. Microaggregation is an alternative method for generalization for different kinds of attributes like categorical, nominal, ordinal and for continuous attributes. Microaggregation is a pertubrative method such that the computed statistics of original dataset do not differ statistically from pertubrated dataset. It is generally applied on quasi- identifiers for de-identification in two basic steps, partition and aggregation [10]. Initial step is used for partitioning the original data records into several clusters. Each cluster should have equal number of records. In aggregation, mean or median is used as aggregate operators which minimize the information loss.

b) **Identify the sensitive attributes**

Sensitive attributes are anonimized by using t-closeness with earth mover's distance (EMD) and micro aggregation [4]. Here, EMD is used to measure the transforming cost of one distribution say A into another distribution say B by moving the probability mass and declared as EMD (A,B). For numerical attributes the distance between two binaries (numerical values) is based on the number of binaries between them. If the numerical attribute takes the values {u1, u2, u3,…., un}, where um < un if m < n, then ordered distance(um, un)=| m - n |/(p-1). Now, if A and B are distributions over {u1, u2, u3, ……, un} then assign probabilities am and bm to um, and the EMD for the ordered distance can be calculated as:

$$M(,) = \frac{1}{\sum_{p-1=1=1}^{p} |\sum -|}$$

c) **Identify the insensitive attributes**

Insensitive Attributes like name, country names etc., are generally not neecessary to anonymize.

d) **Identify the quasi-identifiers**

The main focus is on quasi identifiers. It is generally used for de-anonymization by the adversary who already has background knowledge. Identifying the quasi identifiers that are of interest and anonymization is done on the structural data at different levels.

e) Interpretation: By using t-closeness we can ensure that privacy is preserved to the maximum extent when compared with l-diversity and k-anonymity. In proposed system, the quasi-identifiers are considered as sensitive attributes because these identifiers may lead to de- identification of other attributes. Level wise anonymization of quasi identifier i.e., zip code is shown in Figure 3.

Results

The cardinality of an equivalence class is measured using Discernibility Metric (DM). The penalty is assigned to each tuple using DM. It is based on how many tuples in the transformed dataset are indistinguishable from original data set. Let 'a' be a tuple from the original table A, and let GA∗ (a) be the set of tuples in an anonymize table A∗ indistinguishable from t or the set of tuple in A∗ equivalent to the anonymized value of t. Then, DM is defined as follows:

$$DM(A^*) = \sum_{a \in A} |GA^*(a)|$$

Here, the Table 1 shows that the number of Quasi- identifiers including risk % for different thresholds. Risk % is minimum for minimum threshold. With increase in threshold risk % also get increased to a maximum extent and after reaching maximum threshold there is no change in risk % (Figure 4).

Here, risk analysis is performed on quasi identifiers. By comparing the risk percentage for the given input and varying the threshold, it is observed that for minimum threshold 'T' the risk percentage is low and risk cannot be minimized i.e., percentage of risk is stabilized for a particular threshold.

Sum of Squared Error (SSE) is used to calculate Accuracy. SSE is defined as follows:

$$SSE=\Sigma=1\ \Sigma_{y \in Ck}\ (ck, y)^2$$

Here, *distance* is the standard Euclidean (L2) distance in Eucildean

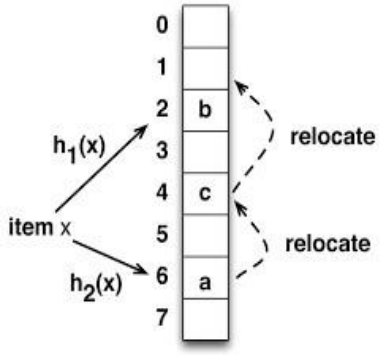

Figure 2: For inserting the value x.

Level-0	*Level-1*	*Level-2*	*Level-3*	*Level-4*
52784	5278	527	52	5
53434	5343	534	53	5
53737	5373	537	53	5
54281	5428	542	54	5
54321	5432	543	54	5

Figure 3: Level based Anonymization for Zip codes.

space between two objects. The centroid that minimizes the SSE is mean for the cluster. Entropy is a function used to calculate "Information Gain/Loss" that satisfies. Information 1 and 2 (m1m2)

Information 1 (m1)+Information 2 (m2). Where m1m2 is the probability of event 1 and event 2, m1 is the probability of an event1 and m2 is the probability of an event 2. Mathematics - Logarithm Function (log)

Informationy(my)=log2(my) I(Y)=log_2(m_y)

Entropy=H(Y)=E(I(Y))=Σy px I(y)=- Σy px log2 px

H stands for entropy and E for Ensemble. The entropy of a distribution with finite domain is maximized when all points have equal probability. Bigger is the entropy, more is the event unpredicatble. Higher entropy mean that the events being measured are less predictable. 100% predictability=0 entropy. Information gain is positive when there is a decrease in entropy from choosing classifier/representation. A decrease in entropy signifies decrease in unpredictability, which also means an increase in predictability (Table 1).

For different classes, accuracy and gain percentage is calculated as shown in Table 2 and the results clearly shows that there is no gain/loss percentage for l-diversity as shown in Figure 5. Results are analyzed for different classes and showed that t-closeness method yeilds maximum gain/loss percentage for minimum threshold t=0.009 as shown in Figure 6.

When threshold reaches the stabilization point increase in threshold does not effect Gain/Loss percentage and Accuracy. By comparing l-diversity with t-closeness accuracy percentage is increased to a greater extent and gain/loss percentage is maximum for minimun threshold in t-closeness method.

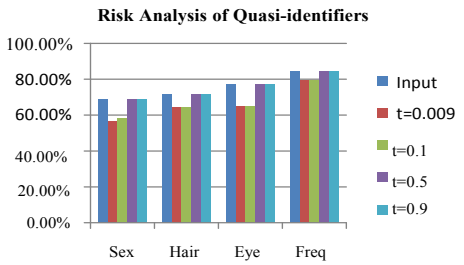

Figure 4: Risk Analysis.

Quasi-Identifier	INPUT	t=0.009	t=0.1	t=0.5	t=0.9
Sex	68.91%	56.68%	58.10%	68.91%	68.91%
Hair	71.68%	64.45%	64.44%	71.68%	71.68%
Eye	77.13%	65.10%	65.10%	77.13%	77.13%
Frequency	84.45%	79.34%	79.33%	84.45%	84.45%

Table I: Risk Analysis.

Class	l-diversity				t-closeness (t=0.9)			
	Accuracy%		Gain/Loss%		Accuracy%		Gain/Loss%	
	Input	Output	Input	Output	Input	Output	Input	Output
Hair	18.5	18.5	18.5	0	18.5	18.5	18.5	0
Eye	22.2	22.2	11.1	0	14.8	30	3.7	75
Sex	77.7	77.7	48.2	0	48.2	100	18.5	108

Table 2: Classification accuracy for l-diversity and t-closeness.

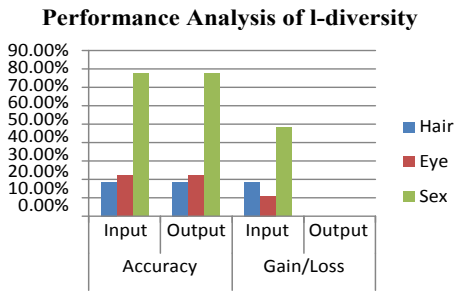

Figure 5: Performance Analysis of l-diversity.

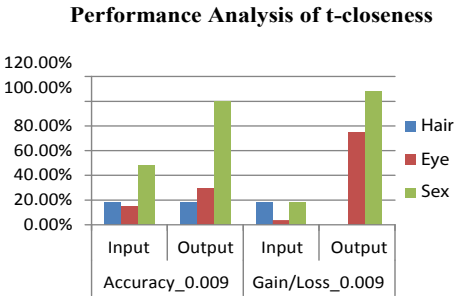

Figure 6: Performance Analysis of t-closeness.

Conclusion

In this paper data anonymization is performed on structural data using cuckoo filters and t-closeness. Set membership test is performed to check whether a particular user is a member of a group or not using Cuckoo filters. And then t-closeness is applied on sensitive attributes and quasi-identifiers to find out the level of de-anonymization. Results have shown that risk percentage of quasi-identifiers is less for minimum threshold. And with t-closeness accuracy percentage increased to a greater extent when compared with l-diversity and gain/loss percentage is high for minimum threshold. The proposed system can be further enhanced for non-structural data.

References

1. Benjamin CMF, Ke W, Rui C, Philip SY (2010) Privacy-Preserving Data Publishing: A survey of Recent Developments, ACM Surveys 42: 4-14.
2. Wei P, Feng L, Xukai Z, Jie W (2014) A Two-Stage Deanonymization Attack against Anonimized Social Networks, IEEE Transactions on Computers 63: 2.
3. Gilbert W, Thorsen H, Engin K., Christopher K (2010) A Practical Attack to De-anonymize Social Network Users, IEEE Symposium on Security and Privacy, pp: 223-238.
4. Bloom BH (1970) Space/time trade-offs in has coding with allowable errors. Communications of the ACM 13: 422-426.
5. Bin F, David GA, Michael K, Michael DM (2014) Cuckoo Filter: Practically Better Than Bloom.
6. Xiaozhou L, David GA, Michale K, Michale JF (2014) Algorithimic Improvements for Fast Concurrent Cuckoo Hashing, EuroSys'14, Amsterdam, Netherlands ACM 978-1-4503-2704-6/14/04.
7. Ciriani V, Capitani S, Foresti S, Samarati P (2007) K-Anonymity, Springer US Advances in Information Security.
8. Ashwin M, Johannes G, Daniel K, Muthuramakrishnan V (2007) L-Diversity: Privacy beyond k-Anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), Volume Issue 1.

9. Tripathy B, Maity A, Ranajit B, Chowdhuri D (2014) A fast p-sensitive ldiversity Anonymisation algorithm, Recent Advances in Intelligent Computational System (RAICS). 2011 IEEE 741-744.

10. Ninghui L, Tiancheng L, Suresh V (2007) T-closeness: Privacy Beyond k-Anonymity and l-Diversity, IEEE 23rd International Conference on Data Engineering (ICDE) 106-115.

11. Jordi SC, Josep DF, David S, Sergio M (2015) T-closeness through Microaggregation:Strict Privacy with Enhanced Utility Preservation. IEEE Transactions on Knowledge and Data Engineering.

12. Domingo FJ, Soria CJ (2015) T-closeness to Differential Privacy and vice versa in Data Anonymization. Knowledge-Based System, pp: 151-158.

13. Ninghui L, Tiancheng L, Suresh V (2010) T-closeness: A new privacy measure for data publishing. IEEE Transactions on knowledge and Data Engineering, pp: 943-956.

14. Josep DF, Vicenc T (2005) Ordinal, Continuous and Heterogeneous k-Anonymity through Microaggregation. Data Mining and Knowledge Discovery, Springer Science + Business Media, Inc.