PAPER Special Section on Security, Privacy and Anonymity of Internet of Things

Achieving High Data Utility K-Anonymization Using Similarity-Based Clustering Model

Mohammad Rasool SARRAFI AGHDAM[†a)], *Nonmember and* **Noboru SONEHARA**[†,††], *Fellow*

SUMMARY In data sharing privacy has become one of the main concerns particularly when sharing datasets involving individuals contain private sensitive information. A model that is widely used to protect the privacy of individuals in publishing micro-data is k-anonymity. It reduces the linking confidence between private sensitive information and specific individual by generalizing the identifier attributes of each individual into at least *k-1* others in dataset. K-anonymity can also be defined as clustering with constrain of minimum *k* tuples in each group. However, the accuracy of the data in k-anonymous dataset decreases due to huge information loss through generalization and suppression. Also most of the current approaches are designed for numerical continuous attributes and for categorical attributes they do not perform efficiently and depend on attributes hierarchical taxonomies, which often do not exist. In this paper we propose a new model for k-anonymization, which is called Similarity-Based Clustering (SBC). It is based on clustering and it measures similarity and calculates distances between tuples containing numerical and categorical attributes without hierarchical taxonomies. Based on this model a bottom up greedy algorithm is proposed. Our extensive study on two real datasets shows that the proposed algorithm in comparison with existing well-known algorithms offers much higher data utility and reduces the information loss significantly. Data utility is maintained above 80% in a wide range of k values.

key words: *anonymization, privacy preserving data mining, K-anonymity, algorithm*

1. Introduction

Nowadays in communication, smart devices, social networks and Big Data era, privacy has become one of the main concerns of individuals and data publishers. In general there are many service providers and organizations such as hospitals, which collect and store huge amount of information on individuals as a process of their common operations. The collected micro-data (individual level data) contains quasi-identifier (QID) attributes and private sensitive attributes. QID attributes are type of attributes, which are used to identify an individual. For instance, Date of Birth (Age), Zip Code and Sex are QID attributes. On the other hand private sensitive attributes, for example Disease or Salary, are type of attributes that are not normally shared with public [1]. Organizations such as hospitals are required to share or release the collected information for the purpose of research and detailed data analysis. For example the medical history of patients as shown in Fig. 1 (a) may be released by a hospital to assist medical studies. Publishing the collected micro-data for research would be very helpful for researchers to investigate the correlation between different attributes such as the relation between a certain disease and location. However, publishing the collected data containing private sensitive information (Disease) would bring up some privacy concerns. Due to the existence of private sensitive attribute such as Disease in the dataset the data publisher must ensure that no adversaries can identify the disease of any patient accurately.

Although the identifying information such as name and social security number are removed before releasing the data, disclosing the private sensitive information of individuals and re-identifying them uniquely is still very much possible due to the existence of QID attributes in the released data [1], [3], [4].

Based on the previous studies on US population using 2000 census data [2], disclosing an individual's full Date of Birth, Sex and Zip Code allows for unique identification of 63% of the US population. This clearly presents the high possibility of re-identification once one's information including private sensitive and QID attributes are shared with the third party and also it points out the main reason of privacy concerns in data publishing [2]. Hence for exercising data analysis and data mining while protecting privacy of individual privacy preserving data mining concept was introduced [6].

Tuple	**Age**	**Sex**	**Zip Code**	**Disease**
T1	**35**	**Male**	**2370**	Fever
T2	45	Male	2410	HIV
T3	20	Female	5490	Headache
T4	25	Male	5410	Cancer

(a) Patient Dataset Released by Hospital

Name	**Age**	**Sex**	**Zip Code**
Ross	35	Male	7415
Rachel	40	Female	7942
Joey	**35**	**Male**	**2370**
Monica	35	Female	6578

(b) Voter Registration Dataset

Fig. 1 Sample of linking attack between (a) released dataset by hospital and (b) external dataset

Manuscript received December 13, 2015.
Manuscript revised April 17, 2016.
Manuscript publicized May 31, 2016.

[†]The authors are with the Graduate University for Advanced Studies (SOKENDAI), School of Multidisciplinary, Informatics Department, Tokyo, 101–8430 Japan.

[††]The author is with National Institute of Informatics (NII), Tokyo, 101–8430 Japan.

a) E-mail: rasool@nii.ac.jp

DOI: 10.1587/transinf.2015INP0019

There are a lot of external data sources, which are available on Internet and accessible to everyone such as Voter Registration list. QID attributes which are in common between external dataset (with identifiers) and released dataset by hospital (with private sensitive information) can be joined and establish a link. This link between these two datasets could associate private sensitive information to individuals and result in re-identifying individuals uniquely, which discloses their private sensitive information. Technically this is known as "linking attack" [1], [3] and [4].

For instance, linking attack between the Patient dataset released by a hospital and Voter Registration dataset as an external dataset is shown in Fig. 1. In this linking attack, the disease of Joey is identified accurately by an intruder therefore it can be concluded that Joey's privacy is violated.

Thus to protect the privacy of individuals against the possible re-identification k-anonymity was proposed by Samarati and Sweeney [1], [3], [4]. K-anonymity suggests modifying the values of QID attributes through generalization and suppression so that each record in the released dataset is indistinguishable from at least *k-1* other records within the same dataset along QID attributes. K-anonymity reduces the linking confidence between k-anonymous released dataset and the external dataset by *1/k* ratio and protects the privacy of individuals to some extent. *K* value in k-anonymity is the anonymization degree and it is selected based on the desired level of privacy.

In Fig. 2, the effect of applying k-anonymity model on the Patient dataset shown in Fig. 1 (a) is illustrated. The linking confidence is reduced by the ratio of 1/2 and the exact identification of Joey as an individual and his specific disease is now impossible.

In k-anonymity model generalization and suppression are the two main methods. Both methods are technically recoding the values of QID attributes in original dataset. In generalization, the original values of QID attributes are replaced by more general values such as intervals or set of distinct values. For instance in attribute Age, the value 25 could be replaced by [20~30] and for attribute Sex, Male could be replaced by Person or [Male, Female]. Suppression can be defined as specific type of recoding in which the values of data record in original dataset is recoded to null values [1], [3] and [4].

By generalizing or suppressing original data records to form k-anonymous dataset, some information loss occurs.

Tuple	**Age**	**Sex**	**Zip Code**	**Disease**
T1	35~45	Male	2***	Fever
T2	35~45	Male	2***	HIV
T3	20~25	Male, Female	54**	Headache
T4	20~25	Male, Female	54**	Cancer

Fig. 2 Effect of K-anonymization on patient dataset released by hospital (Fig. 1 (a)), 2-anonymous patient dataset

Information loss in k-anonymity model is an unfortunate and inevitable consequence [4]. The information loss due to the distortion on QID attributes through generalization or suppression reduces the utility of anonymized-data. It makes the anonymized-data to be less accurate and accordingly less useful for specifically researchers or data miners. Thus, one of the main challenges in k-anonymization is to minimize the information loss while ensuring that the released data is k-anonymous. So high utility anonymized-data could be obtained while the privacy of individuals is also protected [4]. However, there is a tradeoff relationship between the privacy level and the quality of anonymized-data and due to this tradeoff, performing anonymization with maximum privacy and obtaining maximum utility for anonymized-data is not possible. Moreover, the issue of finding a k-anonymization with minimal information loss is also proven to be NP hard problem [5], [7], [8]. Therefore, heuristic algorithms could be one of the possible approaches to minimize high information loss problem in k-anonymization [9]–[11] and [12].

Most of real world datasets contain both numerical and categorical attributes. As a matter of fact most of QID attributes in micro-data are assumed to be categorical type attributes [14]. Most of previous approaches on k-anonymization have considered mainly numerical attributes and regarding categorical attributes they depend on extra information such as hierarchical taxonomies, which often does not exist in real life applications [19].

In this work we focused on information loss and data quality issue in k-anonymization. The main previous works and well-known algorithms are reviewed carefully. We introduce some of the information quality metrics in order to calculate the information loss and data utility. Then, we propose our new model on k-anonymity and define the similarity measurement for categorical attributes and distance calculation for numerical and categorical attributes. We also introduce a greedy algorithm with bottom-up approach based on the proposed model. Finally, we evaluate the proposed algorithm, regarding information loss and utility of anonymized-data, and compare it with other existing well-known algorithms.

2. Basic Definitions

Considering the original dataset T which contains the information on each individual in n attributes {A1, . . . , An} the main terminologies are defined as below.

Quasi-Identifier attributes: set of attributes in dataset T that can potentially join with external datasets to reveal private information of individuals. For example Age, Sex and Zip Code attributes in Fig. 1 (a) are quasi-identifiers, which can establish a link between Patient dataset and Voter Registration dataset.

Equivalent class: An equivalent class E of dataset T is a set of all tuples in T containing identical values with respect to QID attributes. For instant T1 (tuple 1) and T2 (tuple 2) in Fig. 2 form an equivalent class (E1) with respect

to attributes Age, Sex and Zip Code.

K-anonymity: A dataset T is said to be k- anonymous with respect to the QID attributes if the size of every equivalent class is greater or equal to pre-defined *k* value.

3. Existing Techniques

K-anonymity is achieved through generalization and suppression where original values are replaced with more generalized values. Typically numerical attributes are generalized into intervals and categorical attributes are generalized into a set of distinct values or in case hierarchical taxonomy for attributes exist, a single value that represents such a set. Generally there are three types of generalization models, 1) global recoding, 2) multidimensional recoding and 3) local recoding.

In global recoding the values in original dataset is generalized at the domain level. Therefore if a lower level domain needs to be generalized to the higher domain, all the values in the lower level domain are generalized to the higher domain. There are a lot of works, which are based on global recoding generalization such as [4], [12], [13], [16], [18]. One of the global recoding generalization methods is Incognito [15]. Incognito produces minimal full domain generalization. Domain level generalization causes overgeneralization of original dataset, which results in very high information loss.

In multidimensional recoding and local recoding, the generalization is taking place at cell levels [7], [8], [10], [11]. They do not cause overgeneralization, which lead to more flexible generalization with possibility of causing less information loss. Multidimensional recoding problem is studied in [17] which suggests an efficient partitioning method for multidimensional recoding anonymization. Mondrian is a heuristic algorithm with top-down approach. It considers that the data are sorted along all attributes. It starts from the whole dataset as a single group and splits the group into segments considering that the minimum allowed group size is *k* [17]. However it is not practical in most of cases involving categorical attributes because this method requires the total order for each attribute and in categorical attributes there is no meaningful order.

The work in [19] introduces utility-based anonymization through local recoding generalization. It introduces a new quality metric that calculates the information loss due to generalization for both numerical and categorical attributes and actually uses this quality metric for clustering the tuples. However this method depends on hierarchical structure regarding categorical attributes and it assumes that for every categorical attribute in the dataset the hierarchical taxonomy is defined and exists, which is not so realistic considering real life applications.

4. Information Loss Metrics

In anonymization process original data will be distorted. In order to measure how useful the anonymized-data might be to the data users, various information loss metrics have been proposed. For instance, Normalized Certainty Penalty (NCP) [19] defines information loss due to generalization for both numerical and categorical attributes. For numerical attributes the NCP of a cell on numerical attribute A_i that lays on equivalent class G is defined as shown below.

$$NCP_{A_i}(G) = \frac{Max^G_{A_i} - Min^G_{A_i}}{Max_{A_i} - Min_{A_i}} \quad (1)$$

In case of categorical attributes, the NCP of the equivalent class G in A_i attribute is defined as follows.

$$NCP_{A_i}(G) = \begin{cases} 0, & Card(u) = 1 \\ Card(u)/Card(D_i), & Otherwise \end{cases} \quad (2)$$

Where, *Card(u)* is the number of distinct values of A_i in G and *Card(D_i)* is the total number of distinct values of attribute A_i. In NCP the information loss due to suppression is considered to be maximum, equal to one. Information loss and utility components oppose each other. Original dataset has no information loss, 0%, therefore its utility is considered to be 100%. However, anonymized-dataset has some information loss therefore its utility is going to be lower than utility of original dataset. By normalizing the total NCP between zero and one, the utility of anonymized-data could be defined as a function of information loss (total NCP) as follows [21].

$$Utility = 1 - NCP_{Total} \quad where, \ 0 \le NCP_{Total} \le 1 \quad (3)$$

The *ILoss* metric proposed in [20] calculates the information loss of a specific value of a record, which is generalized. *ILoss* metric is expressed as follows.

$$ILoss(v_g) = \frac{|v_g| - 1}{|D_A|} \quad (4)$$

In this expression $|v_g|$ is the number of domain values that are descendants of v_g and $|D_A|$ is the number of domain values in the attribute A of v_g and this metric requires all original data values to be at the leaves in the taxonomy.

The Classification Metric CM [12], charges a penalty for a record if its private value differs from the majority of the private values in its group or if the record is totally suppressed.

Minimal Distortion (MD) [16] is a single attribute measure and it defines the information loss as number of instances, which are made indistinguishable. For example if ten records are generalized in Sex attributes from "Male" or "Female" to "Person", the information loss is equal to ten.

The Discernibility Metric (DM) [13] assigns penalty to each record based on the number of records indistinguishable from that record in anonymized table. The DM metric defines information loss for generalization and suppression, which can be expressed mathematically as follows.

$$C_{DM}(g, k) = \sum_{\forall Es.t.|E \ge k|} |E|^2 + \sum_{\forall Es.t.|E<k|} |D|\,|E| \quad (5)$$

In this expression E is the equivalent class and |D| is the size

of the original dataset. The first sum calculates the information loss for generalized tuples and the second sum computes the information loss due to suppression. The information loss in both MD and DM is defined based the size of the group that the record is generalized and even though the DM is more accurate than MD, in k-anonymization methods which are near optimum the size of the groups are close to k value which makes these metrics less practicable.

For information loss measurement and evaluation in this work, Normalized Certainty Penalty (NCP) [19] is considered as it is defined for both numerical and categorical attributes and it calculates information loss and data utility due to generalization and suppression accurately.

5. Proposed Similarity-Based Clustering Model

As it was mentioned earlier, the main issue in k-anonymity is the huge information loss that occurs due to distortion of original values in anonymization process. Also it was mentioned that real datasets are consist of numerical and categorical attributes and current approaches are mainly consider the numerical data or if they consider categorical attributes they require and depend on additional information such as attribute hierarchical taxonomies which mostly do not exist in real life applications.

In our approach k-anonymity problem is defined as a clustering issue. For clustering a dataset the distances between tuples, which represents information loss, are calculated based on new similarity measurement for categorical attributes. After clustering the dataset with respect to k value, all clusters are anonymized through local recoding.

A dataset is called k-anonymous dataset when for every record in the dataset there are at least *k-1* other records identical to it along the quasi-identifier attributes. K-anonymity could also be defined from clustering point of view.

Definition 1. K-anonymity is clustering original dataset T with constrains of minimum of k tuples (data records) in each cluster.

The main challenge of clustering approach in k-anonymization is slightly different than common clustering problem. Typically in clustering number of clusters in the dataset is important however in k-anonymity number of data records (tuples) in each cluster is essential.

The main goal in clustering approach in anonymization is to find the k closest tuple in dataset and group them all together. So in each cluster there are at least k tuples, which satisfies k value condition in k-anonymity and all tuples in the same cluster have minimum possible distance from each other thus the information loss in each cluster is minimized.

Due to the existence of categorical attributes the distance between tuples cannot be simply calculated. Therefore in this work we introduce a new similarity measurement for categorical attributes and calculate the distance based on the measured similarities. By having the distances in categorical and numerical attributes the total distance is calculated and clustering can be performed.

5.1 Similarity Measurement and Distance Calculation for Categorical Attributes

Regarding categorical attributes, distance is not well defined due to the nature of categorical attributes and the problem of representing the values in categorical attributes numerically. In some previous works the distance in categorical attributes is defined with the help of the hierarchical taxonomy [e.g., 19]. However, the hierarchies often do not exist or defined in real life applications. In Similarity-Based Clustering (SBC) model the distance between the values in categorical attributes is defined based on the context and the observation probability of values in each attribute. It is efficient and easily adjustable depending on the number of categorical attributes.

The first step in this approach is to construct the contingency table. The contingency table measures the observation probability for each value of categorical attribute A_j and assesses the similarity between y_1, the value of the first tuple (t_1) in A_j, and the rest of the values in other tuples of A_j. By knowing which values in A_j has the most and least similarity to y_1 the distances between t_1 and the rest of the tuples in A_j could be defined. For instant, let's consider sorted dataset T with total twenty tuples shown in Fig. 3.

There are two categorical attributes, Sex = {Male, Female} and Nationality = {Japan, USA, Iran} as shown in Fig. 3 (a). The attributes are arranged with respect to cardinality order and the contingency table for categorical attributes in dataset T is constructed and shown in Fig. 3 (b). As it is shown below in the contingency table the attribute with higher cardinality that the similarity measurement between its values are going to take place is placed horizontally and the attribute with lower cardinality is placed in the left side of the table vertically.

There is no need to measure the similarity between the values in the attributes with cardinality less or equal to two ($Card(Att_i) \leq 2$). Because for attributes with cardinality equal to one there is only one value and the distance between the identical values is defined as zero. For attributes with cardinality equal to two, there is only one distance to be defined and the distance is defined as maximum, which is equal to one. As shown in Fig. 3 (a) the values of Sex

Tuple	Sex	Nationality
T1	Male	Japan
⋮	⋮	⋮
T20	Female	Iran

(a) Categorical Attributes in Original Dataset T

	Japan	USA	Iran
Male	**4**	4	1
Female	4	1	6

(b) Contingency Table of Dataset

Fig. 3 Categorical attributes in dataset T and (b) its contingency table

Table 1 Contingency table of dataset T and the total number of tuples in each row

	Japan	**US**	**Iran**	**Total No. Tuples**
Male	**4**	4	1	4+4+1 = **9 ≥ k=3**
Female	4	1	6	4+1+6 = 11 ≥ k=3

Table 2 Contingency table of dataset T and the total number of tuples in each row

	Japan	**USA**	**Iran**	**Total No. Tuples**
Male, Female	**8**	5	7	8+5+7 = **20 ≥ k=10**

and Nationality attributes in t_1 (tuple one) are {Male} and {Japan}. By indicating the values in t_1 we start the similarity measurement for attribute with minimum cardinality more than two, which in this example is Nationality attribute.

Also, we need to calculate the total number of tuples in each row of the contingency table as shown in Table 1. If it is greater than or equal to the pre-defined k value then the similarities are measured with respect to the total number of tuples in that row only, else other rows in that specific attribute needs to be considered regarding similarity measurement.

In this example k value is considered to be $k = 3$ and the total number of tuples in Male row in Table 1 is greater than k value. However if it were not, the Female row also would be considered for similarity measurement between the values in Nationality attributes. The modified contingency table in case of k value $k = 10$ is shown in Table 2.

The main reason for such confirmation on total number of tuples is if k value $k = 10$ is considered no matter how we try, the tuples which are grouped and clustered together are going to be a mixture of Male and Female regarding Sex attribute as there are not enough tuples (more than or equal to 10) with only Male value in their Sex attribute.

Definition 2. Considering a dataset T with two categorical attributes $M = \{m_1, \cdots, m_i\}$ and $N = \{n_1, \cdots, n_j\}$, the probability of observation for each value in attribute N when $i < j$, $1 \leq K \leq i$, $1 \leq L \leq j$ and the total number of tuples in m_K is more than k value, is defined as:

$$P(n_L)_{m_k} = \frac{(|n_L|)_{m_K}}{(|n_1| + \cdots + |n_j|)_{m_K}} \tag{6}$$

The notation $(|n_L|)_{m_K}$ indicates the number of tuples with value of n_L in N and value of m_K in M attribute and $(|n_1| + \cdots + |n_j|)_{m_K}$ means the total number of tuples in attribute N which have the value of m_K. The expression (6) can be expanded for multiple categorical attributes with multiple values.

By calculating all the observation probabilities for each value in attribute $N = \{n_1, \cdots, n_j\}$ and obtaining $P(n_1)_{m_K}, \cdots, P(n_j)_{m_K}$, the similarity between the value of t_1 in attribute N and other values in N could be defined. The closer the $P(n_L)_{m_k}$ is to $P(n_1)_{m_k}$, the more similar n_L is to n_1.

The similarity between the values in Nationality attribute in Table 1 is calculated as shown below.

$$P(\text{Japan})_{\text{Male}} = \frac{(|\text{Japan}|)_{\text{Male}}}{(|\text{Japan}| + |\text{USA}| + |\text{Iran}|)_{\text{Male}}} = \frac{4}{9}$$

$$P(\text{USA})_{\text{Male}} = \frac{(|\text{USA}|)_{\text{Male}}}{(|\text{Japan}| + |\text{USA}| + |\text{Iran}|)_{\text{Male}}} = \frac{4}{9}$$

$$P(\text{Iran})_{\text{Male}} = \frac{(|\text{Iran}|)_{\text{Male}}}{(|\text{Japan}| + |\text{USA}| + |\text{Iran}|)_{\text{Male}}} = \frac{1}{9}$$

Therefore since $P(\text{Japan})_{\text{Male}}$ is closer to $P(\text{USA})_{\text{Male}}$ than $P(\text{Iran})_{\text{Male}}$ then Japan is more similar to USA and less similar to Iran. Therefore the similarity order for Nationality attribute is defined as {Japan, USA, Iran}.

After measuring the similarity between the values of all categorical attributes in dataset, the distances between the values can be defined. We start with lowest cardinality attribute to highest and distances are defined with respect to the measured similarities from least similarity to most. In this example Sex attribute with cardinality two is the lowest and since there is only one distance to be defined (distance between Male, Female) it is defined as maximum distance, $D(\text{Male}, \text{Female}) = 1$. For the second minimum cardinality attribute, which is Nationality in this example, the least similarity is between Japan and Iran and Japan and USA are the most similar values. Distances are defined with respect to the similarity order {Japan, USA, Iran} as shown below.

$$D(\text{Japan}, \text{Japan}) = 0$$

$$D(\text{Japan}, \text{USA}) = \frac{\text{Index of USA in Similarity Order}}{|\text{Card(Nationality)}| - 1} = \frac{1}{2}$$

$$D(\text{Japan}, \text{Iran}) = \frac{\text{Index of Iran in Similarity Order}}{|\text{Card(Nationality)}| - 1} = \frac{2}{2}$$

As shown above, all the distance between values in Nationality attribute is calculated. The numerator is the index of the value in the similarity order that was measured and the denominator is the cardinality of attribute minus one, which basically indicates the number of distances, which need to be defined. Therefore all the distances between values are defined between 0 and 1. The most similar values have smaller distance and the most dissimilar values have the highest possible distance, which is equal to 1. By defining the distances using this method, the most similar values in different categorical attribute will have smallest distances to values at t_1.

In our example in this section, finally by having $D(\text{Male}, \text{Female}) = 1$, $D(\text{Japan}, \text{US}) = 1/2$ and $D(\text{Japan}, \text{Iran}) = 1$ defined, the total distance between t_1 and other tuples in dataset T can be calculated as the sum of the $D(\text{Male}, \text{Female})$ and $D(\text{Japan}, \text{US or Iran})$. If a dataset is a combination of numerical and categorical attributes there is a separated process necessary for numerical attributes for distance calculation and normalization.

5.2 Distance Calculation for Numerical Attributes

For numerical attributes the distance measurement is rather conventional. The distance between two tuples t_1 and t_2 with respect to attribute A_i with values of x_1 and x_2 is defined as shown below.

$$\text{Distance}(t_1, t_2)_{A_i} = \frac{|x_1 - x_2|}{R(A_i)} \tag{7}$$

$R(A_i)$ is the range of A_i attribute and it is defined as $R(A_i) = \text{Max}(A_i) - \text{Min}(A_i)$. Based on this, the total distance between t_1 and t_2 for numerical attributes in dataset T is the sum of the $D(t_1, t_2)_{A_i}$ for every A_i, where A_i for $(i = 1, \cdots, n)$ is the numerical QID attribute in dataset T.

5.3 Total Distance Calculation Between Tuples

After calculating all distances between first tuple and the rest of tuples in numerical and categorical attributes and normalizing both separately, the total distance between tuples can be calculated. Considering the original dataset T with the numerical attributes $\{X_1, \cdots, X_m\}$ and categorical attributes $\{Y_1, \cdots, Y_m\}$, the total distance between two tuples t_1 and t_2 is defined as a sum of the normalized distances in numerical and categorical attributes as shown in Eq. (8). Obviously after the addition the total distance (D_T) will be normalized between 0 and 1.

$$D_T(t_1, t_2) = \sum_{i=1,\ldots,m} \Big(D(t_1[X_i], t_2[X_i])\Big) + \sum_{j=1,\ldots,n} \Big(D(t_1[Y_j], t_2[Y_j])\Big) \tag{8}$$

6. Greedy Bottom-Up Algorithm

Based on the proposed model on similarity and distance measurement for clustering in k-anonymization, we introduce a greedy algorithm with bottom-up approach called Similarity-Based Clustering Algorithm (SBCA). In SBCA, every single tuple is considered as a point in the Euclidean space and the dimension of the space is the number of attributes. Then, the original dataset is sorted and the numerical quasi-identifiers are separated from the categorical ones for similarity measurement and distance calculation.

The contingency table for categorical attributes is constructed and after the similarity measurement all the distances are defined. By having all the distances for categorical attributes and using the formula for distance calculation in numerical attributes the total distances between t1 and other tuples are calculated and normalized.

Then in order to find the *k-1* closest tuples to t1, to place in the same equivalent class, the total distance between t1 and the rest of the tuples in dataset T is calculated and t1 and the *k-1* tuple with minimum distances are moved to merge clause and deleted from T. Considering *k* value, the number of tuples in merge clause must be greater or equal to *k*. Therefore if the group size in merge clause is less than *k* then more tuples need to be added to merge clause. Once the number of tuples in merge clause is equal or greater than *k* value the tuples in merge clause are anonymized through local recoding anonymization. Which means, a range from minimum to maximum will replace the numeric values for numerical attributes and values in categorical attributes will be replaced by a set of distinct values. After each equivalent class is made and anonymized, the contingency table will be updated. Therefore similarity is measured again between the values in categorical attributes and the new distances are calculated. This operation repeated until the total tuples in dataset T is none or less than *k* value.

Input: Original dataset T & K-Value
Output: K - anonymous table T'
1: Sort "T" on Min_Cardinality_Att
2: WHILE |dataset T| ≥ K-Value DO {
3: Obtain First Tuple in Sorted "T"
4: FOR Categorical_ Att:
4.1:Contingency table constructed
4.2:K-Value check
4.3:Similarity measurement
4.4:Calculate distances
5: FOR Numerical_ Att:
5.1:Numerical_Att Distance calculation
6: Calculate Total Distance between "first tuple" and the rest of the tuples in T
7: Cluster k-1 closest tuples to First Tuple into Merge clause
8: Anonymize Merge clause through local recoding & DELETE from T & SAVE T'
9: IF |dataset T| < K Value DO Suppression or Add to last cluster
10: Publish K-anonymous table T'

Fig. 4 Pseudo code of similarity-based clustering algorithm (SBCA)

The remaining tuples could be suppressed (removed from the dataset) or could join the already existing equivalent classes with minimum distance. However in most cases the last equivalent class has the highest information loss and the remaining tuples could be added to the lastly created equivalent class. After this process there will be no more tuple left in original dataset T and the k-anonymous dataset can be published. The pseudo code for the SBCA is shown in Fig. 4.

6.1 Complexity Analysis

There is one main primitive operation in SBCA, which is calculating distances between all tuples and selecting k tuples with minimum distance. Actual computational complexity of SBCA depends on selected k value. The range of k value is $2 \leq K \leq n/2$. "n" is number of tuples in the dataset. The maximum number of distances to be calculated to select k tuples for anonymization is $n - 1$. Considering dataset with n tuples the same process has to be performed n/k times.

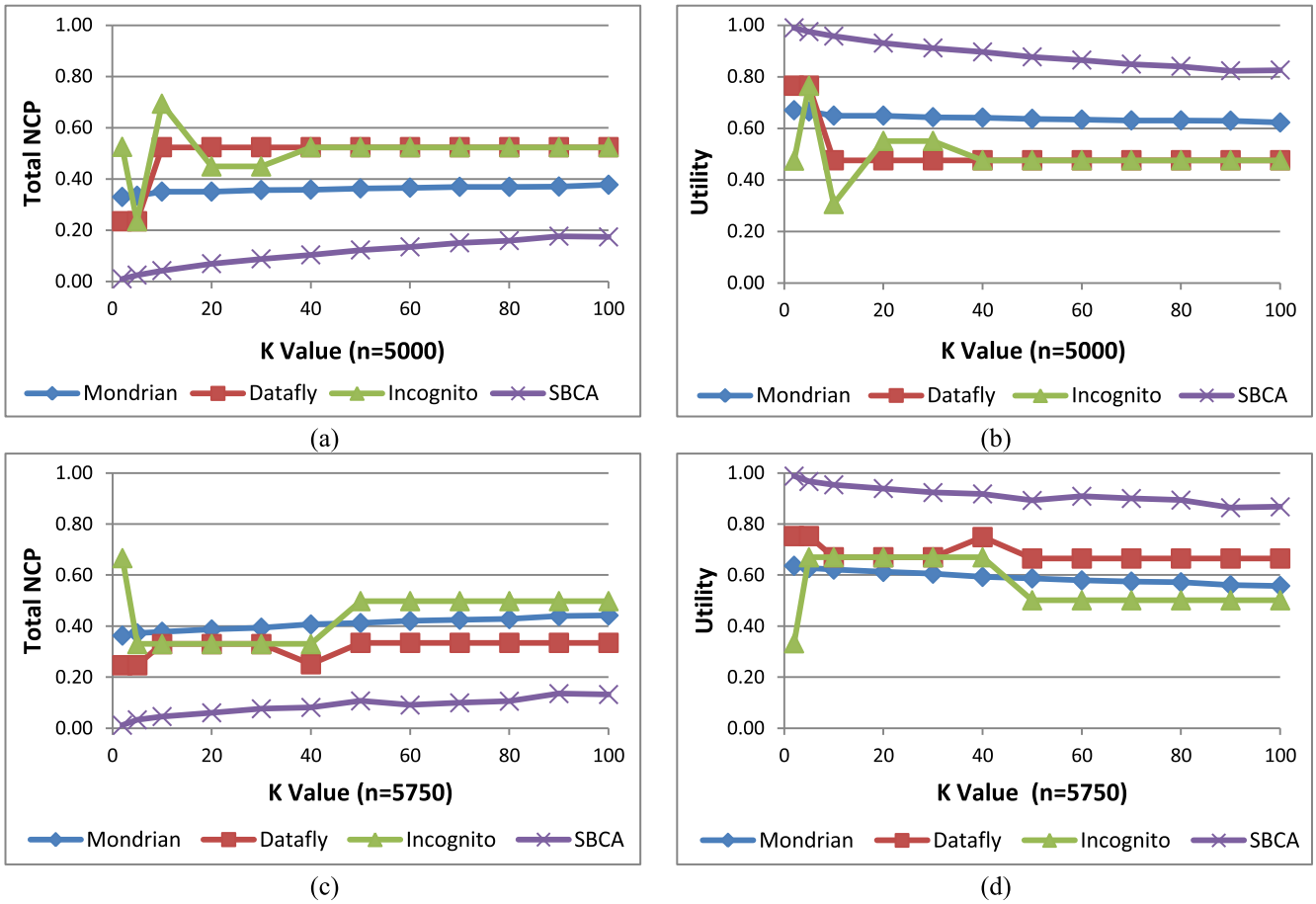

Fig. 5 Information loss and data utility comparison between mondrian, datafly, incognito and SBCA on adult dataset ((a), (b)) and ISP dataset ((c), (d))

Therefore the computational complexity is calculated as $O((n/k) * (n - (n * k/2) - 1))$. The worst-case scenario is when k value is minimum $k = 2$, then the complexity is $O(n^2)$. However in real applications, as the number of records in dataset is increasing, the selected k value for anonymization is increasing relatively, which will lower the computational cost.

7. Empirical Evaluation

In order to examine the performance of SBCA, we calculated the information loss and measured the utility of anonymized-data and compared the results with existing well-known algorithms such as Incognito, Datafly and Mondrian [15]–[17]. The information loss is measured using the total Normalized Certainty Penalty (NCP) metric with maximum value of one and the minimum of zero.

Regarding the sample datasets, we have used the Adult dataset from University of California Irvine (UCI) Machine Learning Repository, which contains census data and has become a benchmark for k-anonymity [22]. The selected data from Adult dataset has 5000 records with 4 different attributes with the distribution over quasi-identifier attributes shown in Fig. 6. The quasi-identifier attributes are Age (cardinality = 67), Sex (cardinality = 2) and native-Country (cardinality = 39). The private sensitive attribute is Salary. N. Corporation as an Internet Service Provider (ISP) has huge storage of data. There was a necessity of publishing some of the collected data, which actually contained private sensitive information about their customers. Therefore it had to be anonymized. As a real case study on k-anonymization through SBCA the dataset was anonymized. This dataset has 5750 data records with 4 attributes. The quasi-identifier attributes are Age (cardinality = 70), Sex (cardinality = 2) and Location (cardinality = 49) with the distribution shown in Fig. 6. The monthly charge of the service is the private sensitive attribute.

The frequency distribution illustrated in Fig. 6 shows that mostly the data is not normally distributed over the quasi-identifier attributes in both datasets. Especially it appears to be long-tailed distribution over Location, Native-Country and Sex attributes.

For the simulation of Mondrian, Incognito and Datafly algorithms we have used University of Texas at Dallas (UTD) anonymization toolbox, which is available online [23].

As it is shown in Fig. 5 (a) and (c) the total NCP of SBCA for the range of k values (k = 2, 5, 10, 20, 30, 40,

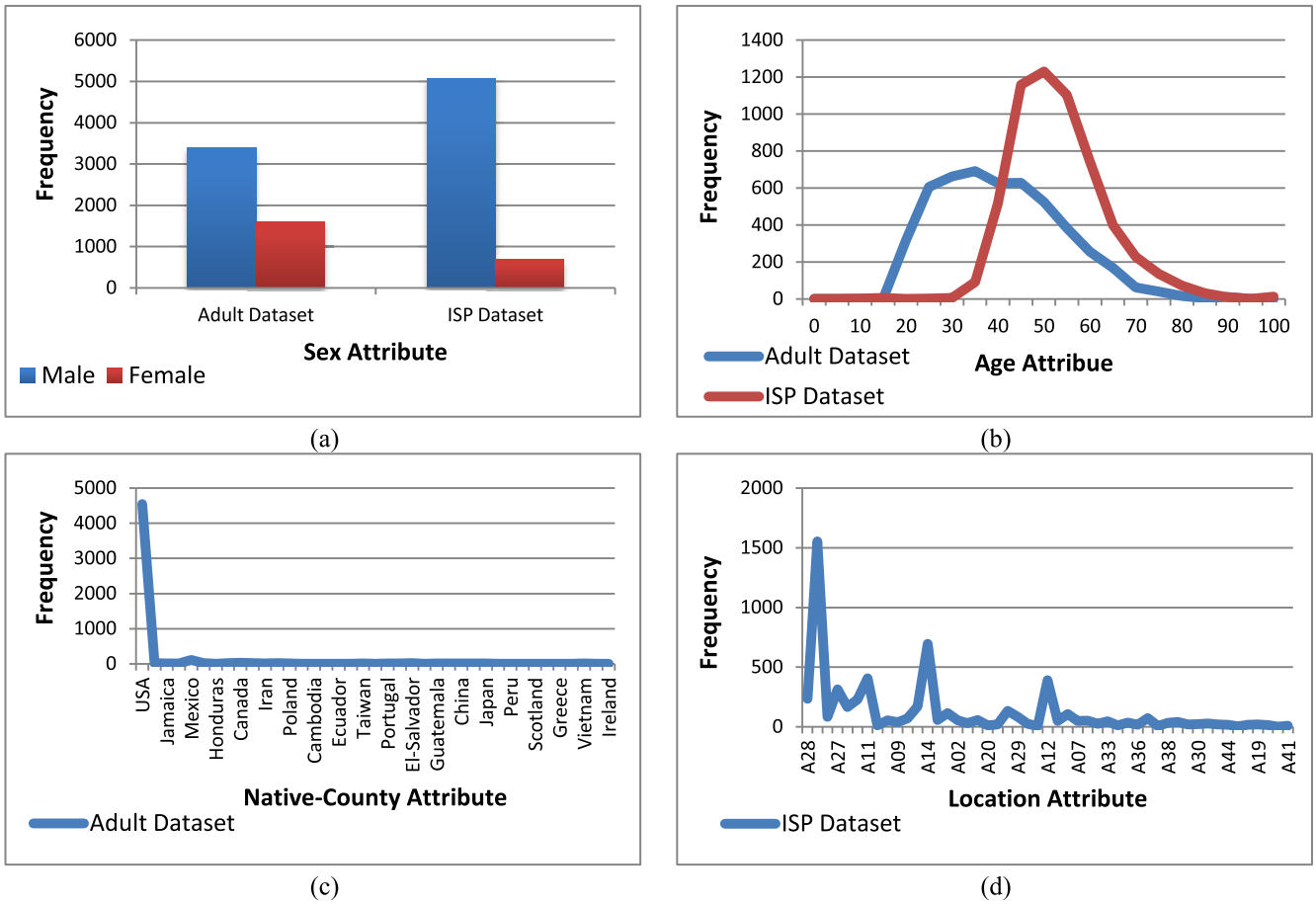

Fig. 6 Frequency distribution of quasi-identifier attributes (a) sex, (b) age, (c) native-country and (d) location in adult and ISP datasets

50, 60, 70, 80, 90, 100) is much less than other algorithms. It is clear that SBCA offers anonymization with much lower information loss by keeping the same privacy level (k value). This advantage in reducing the information loss while maintaining the anonymity level would result in a very high utility anonymized-data. The information loss in both datasets regarding SBCA is maintained below 20%. The comparison on utility of anonymized-data between SBCA, Mondrian, Incognito and Datafly algorithms are also shown in Fig. 5 (b) and (d). As it was expected by having the result on total NCP measurements, the utility of anonymized-data in SBCA is much higher than other algorithms. The data utility in both datasets regarding SBCA is maintained above 80%.

The reason that Total NCP is much lower in SBCA comparing to other algorithms shown in Fig. 5 (a) and (c) especially in lower k value, such as k = 2, is efficient clustering and generalization that is taking place in SBCA. As it was mentioned in Sect. 3, overgeneralization is one of the main reasons of having huge information loss in anonymization.

One of the ways to investigate overgeneralization is by counting number of groups and number of tuples in each group. However, how close the tuples are together in each group is a very crucial point in minimizing information loss. In k-anonymization the minimum number of tuples in each group is specified by k value. Therefore it is desirable that each group (cluster) has exactly k number of tuples.

Table 3 Anonymized-data result analysis at k = 2

Algorithm	Dataset	K-value	No. of Unique Set of Results
Mondrian	Adult	2	209
	ISP		509
Incognito	Adult	2	22
	ISP		49
Datafly	Adult	2	16
	ISP		14
SBCA	Adult	2	582
	ISP		2016

As it was not possible to count number of groups and tuples in each group for each of the algorithms shown in Fig. 5, from the anonymized-data number of unique set of results are counted and shown in Table 3. As it is shown number of unique set of results in SBCA is much higher than other algorithms, which indicates more number of groups. More number of groups indicates less overgeneralization effect and lower information loss.

Domain level anonymization is causing overgeneralization, which is used in Incognito and Datafly algorithm. At k = 2 regarding Incognito, 100% of Sex attribute information is lost. Meaning that in anonymized-data the record sex value is not clear. Regarding Datafly 50% of Age attribute in Adult dataset is lost. In Mondrian almost 98% of Sex attribute information is lost due to the lack of meaningful order in categorical attributes.

Regarding tradeoff relationship between the privacy and utility, in Fig. 5 (c) and (d) tradeoff relationship can be observed clearly for SBCA and to some extent for Mondrian that by increasing the k value, which is the privacy level, the utility is decreasing. However due to overgeneralization effect in Incognito and Datafly which causes high information loss even in small k values, such relation could not be observed.

8. Conclusion and Future Work

In this work, we have studied the information loss issue due to generalization in k-anonymity model. We have reviewed some of the previous works and information loss metrics in this domain. We have also emphasized on the issue that datasets are a combination of numerical and categorical attributes and yet most of the existing models are not designed for categorical attributes or they depend on the hierarchical taxonomies which often do not exist or defined in real life applications. In order to solve the indicated issues we have proposed a new model based on clustering, achieving k-anonymity through local recoding generalization for datasets including numerical and categorical data without hierarchical taxonomy. Then based on the proposed model we have suggested greedy algorithm and compared its simulation results on real datasets to well-known algorithms. The results show that the information loss due to generalization is significantly reduced and it offers much higher utility for anonymized-data in addition of being independent of attributes hierarchical taxonomies.

As this work mostly focuses on achieving high utility anonymization and reducing information loss, for future works, analyzing the scalability of the current model and possibly improving it to anonymize large-scale datasets for Big Data applications using the MapReduce framework is considered.

References

[1] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol.10, no.5, pp.557–570, 2002.
[2] P. Golle, "Revisiting the Uniqueness of Simple demographics in the US Population," Workshop on Privacy in the Electronic Society (WPES), pp.77–80, Alexandria, Virginia, USA, Oct. 30, 2006.
[3] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," Proc. ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems (PODS '98), ACM, p.188, 1998.
[4] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowl. Data Eng., vol.13, no.6, pp.1010–1027, Nov./Dec. 2001.
[5] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Approximation algorithms for k-anonymity," Journal of Privacy Technology, pp.1–18, 2005.
[6] R. Agrawal and R. Srikant, "Privacy–preserving data mining," ACM SIGMOD Record, vol.9, pp.439–450, 2000.
[7] A. Gionis and T. Tassa, "k-Anonymization with minimal loss of information," IEEE Trans. Knowl. Data Eng., vol.21, no.2, pp.206–219, 2009.
[8] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," Proc. ACM Sigmod-Sigact-Sigart Symposium, Pods, pp.223–228, 2004.
[9] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "A framework for efficient data anonymization under privacy and accuracy constraints," ACM Transactions on Database Systems (TODS), vol.34, no.2, Article number 9, June 2009.
[10] A. Gionis, A. Mazza, and T. Tassa, "k-Anonymization revisited," Proc. IEEE Int. Conf. on Data Eng. (ICDE), pp.744–753, 2008.
[11] M.E. Nergiz and C. Clifton, "Thoughts on k-anonymization," Journal of Data and Knowl. Eng., pp.622–645, 2007.
[12] V.S. Iyengar, "Transforming data to satisfy privacy constraints," Proc. Int. Conf. on Knowl. discovery and data mining, pp.279–288, 2002.
[13] R.J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," Proc. IEEE Int. Conf. on Data Eng. (ICDE), pp.217–228, 2005.
[14] L. Willenborg and on e Waal, "Elements of Statistical Disclosure Control," Lecture Notes in Statistics, vol.155, pp.1–37, Springer, 2001.
[15] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Incognito: efficient full-domain k-anonymity," Proc. SIGMOD, pp.49–60, 2005.
[16] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol.10, no.5, pp.571–588, 2002.
[17] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," Proc. IEEE Int. Conf. on Data Eng. (ICDE), p.25, 2006.
[18] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. VLDB, pp.139–150, 2006.
[19] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.W.-C. Fu, "Utility-Based Anonymization Using Local Recoding," Proc. Int. Conf. on Knowl. discovery and data mining (KDD), pp.785–790, 2006.
[20] X. Xiao and Y. Tao, "Personalized privacy preservation," SIGMOD '06 Proceedings of the 2006 ACM SIGMOD international conference on Management of data, pp.229–240, 2006.
[21] M.N. Huda, S. Yamada, and N. Sonehara, "On Enhancing Utility in k-Anonymization," International Journal of Computer Theory and Engineering, vol.4, no.4, pp.527–532, 2012.
[22] A. Frank and A. Asuncion, (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[23] UTD Anonymization Toolbox, http://cs.utdallas.edu/dspl/cgibin/toolbox/index.php

Mohammad Rasool Sarrafi Aghdam received his Bachelor of Science degree in Electronics Engineering from Multimedia University, Malaysia in 2011. After completing one-year internship at France Telecom Orange Labs he started his PhD at The graduate University for Advanced Studies (SOKENDAI). He is currently a PhD candidate and his interests include privacy enhancing technologies, privacy preserving data mining, Big Data and privacy issues, wireless network and indoor localization technologies.

Noboru Sonehara is a Professor of Information and Society Research Division, at National Institute of Informatics from 2004. Previously Project Manager, Content Commerce Project, at NTT Cyber Solutions Laboratories from 2001 to 2004. He received the B.E. degree and the M.E. degree from Shinshu University, Nagano in 1976 and 1978, respectively. He then joined NTT, and has been engaged in R&D of facsimile communication, weather forecasting, content ID, and content commerce systems. He received the Ph.D. degree in 1994. He is a member of IIEE and ITE, and a Fellow of IEICE. He is a Director of Information and Society Research Division from 2006.