Journal of Theoretical and Applied Information Technology
31st July 2016. Vol.89. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SIMPLE AND EFFECTIVE METHOD FOR SELECTING QUASI-IDENTIFIER

[1]AMANI MAHAGOUB OMER, [2] MOHD MURTADHA BIN MOHAMAD

[1]Faculty of Computing, University Technology Malaysia
[2]Faculty of Computing, University Technology Malaysia
E-mail: [1] Amani_ome@hotmail.com, [2] murtadha@utm.my

ABSTRACT

In this paper, a new method to select quasi-identifier (QI) to achieve k-anonymity for protecting privacy is introduced. For this purpose, two algorithms, Selective followed by Decompose algorithm, are proposed. The simulation results show that the proposed algorithm is better. Extensive experimental results on real world data sets confirm efficiency and accuracy of our algorithms.

Keywords: *K-anonymity, Privacy Preserving, Quasi-identifier.*

1. INTRODUCTION

The propagation of data along the Internet and access to fast computers with great memory capacities has increased the intensity of data compiled and disseminated about individuals[1]. Needs of this information is valuable in both research and business. Researcher needs for classification, analysis, statistics and computation. But, sharing and publishing the data may put the respondent's privacy at risk.

Data publishing concerned with, authorized or proper disclosure of information to outside organizations or people [2]. Information should be disclosed only when specifically authorized and solely for the limited use specified. So, data holders need to release a version of its private data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful. The most common approach used to preserving the privacy, is by removing all information that can directly link data items with individuals. This process is referred to anonymization. However, the rest of attribute contain information that can be used to link with other data to infer identity of responders, those type of attribute call quasi identifier for example (age, sex, Zipcode…). Popular example which can uniquely determine about 87% of the population in United States. To overcome the problem of this type of linking via quasi-identifiers, k-anonymity concept was proposed [1].

This study focuses on the data representation and selection of quasi identifier. The current anonymity methods needs to distort a big amount of information during anonymization process which decrease the data utility. Incautious publication of quasi-identifiers will lead to privacy leakage.

On the other hand, data set sometimes contains compact values as one attribute, like zip code and telephone number, those attributes used with other attribute to join the data of an individual so as to infer identity of individual, so our proposed algorithm work with a similar type of attribute to decompose the data into sub attributes.

The existing work addressing formal selection of quasi identifier attribute [3]. This algorithms for finding keys/quasi-identifiers exploit the thought of using random samples to tradeoff between accuracy and space complexity, and can be watched as streaming algorithms. Other study addressing QI problem in [4].demonstrating the role of QI in k anonymity.

2. PROBLEM BACKGROUND

Anonymity is always related to the identification of a user rather than the specification of that user. For instance, a user can be identified through his/her SSN but in the absence of an information source that associates that SSN with a specific identity, the user is still anonymous[5]. Ensuring proper anonymity protection requires the investigation of the following different issues [6]. *Identity disclosure protection*. Identity disclosure occurs whenever it is possible to re identify a user, called respondent, from the released data. Techniques for limiting the possibility of re identifying respondents should therefore be adopted. *Attribute disclosure protection*. Identity

Journal of Theoretical and Applied Information Technology
31st July 2016. Vol.89. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

disclosure protection alone does not guarantee privacy of sensitive information because all the respondents in a group could have the same sensitive information. To overcome this issue, mechanisms that protect sensitive information about respondents should be adopted[7].

Record linkage attack is one of the major channels for violating privacy. To address the problem of record linkage attack, different techniques of statistical disclosure control are employed. One such approach called k-anonymity, works by reducing data across a set of key variables to a set of classes [1] [8] . Other variations of k-anonymity can be found in [7] [9] [10] [11]. In a k-anonymized dataset each record is indistinguishable from at least k-1 other records. Therefore, an attacker cannot link the data records to population units with certainty thus reducing the probability of disclosure. However, preserving privacy through statistical disclosure control techniques leads to loss of a big amount of information to satisfy the privacy requirement. Most of the techniques proposed in literature do not focus on the information loss issues. Rather than privacy and computing time to achieve k-anonymity. The method described in this paper maintains a balance between information loss and privacy. Introduced of formal selection of quasi identifier attribute [3] followed by decomposition algorithm deployed to achieve the balance between information loss and privacy.

3. DATA PRE-PROCESSING AND SELECTIVE ALGORITHM

The first objective is to minimize loss of data during anonymization process by filtering out tuple with missing data, un-known data and duplicate tuple. Into the preprocessing stage. Secondly, we seeks to identify quasi identifier attribute, significant minimal attribute subset and to evaluate its significance in terms of personal identity. The initial investigation was aimed at finding a basic attribute subset that is appropriate to identify the maximum number of tuples in the dataset. Insufficient selection results using randomly attribute subset leads to an attribute investigation to find specific attribute subsets identifying each tuple of the dataset. Figure 1: gives an overview of the experimental procedure for quasi identifier attribute selection.

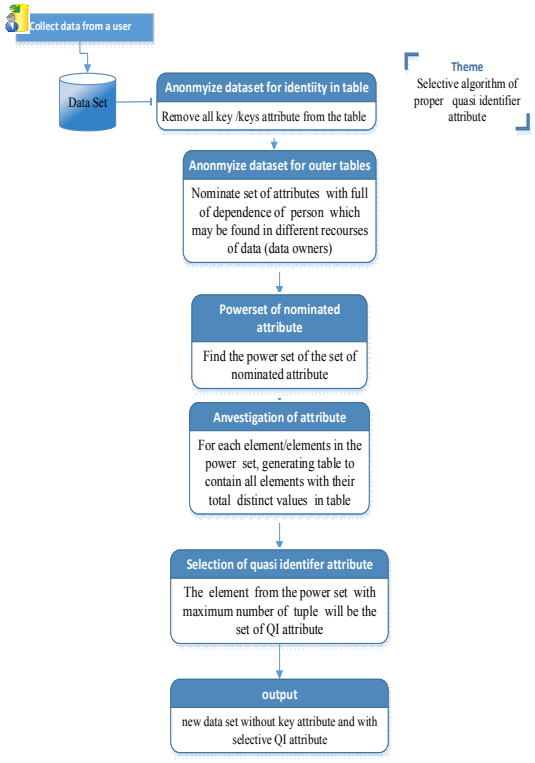

Figure 1: Selective Quasi identifier attributes

To select quasi identifier attribute, firstly we nominate multiple attribute as set, and then we generate P(S) from the set to collect all possible combinations of the attribute. Each element in the set examination by the distinct value in the table, the candidate element from the power set will be the element with maximum distinct value, if the maximum distinct value duplicates in many element, we select the element with minimum attribute.

The intuition is to identify a minimal set of attributes from T that has the ability to (almost) distinctly identify a record and the ability to separate two data records

This study present formal selection procedure depends on probability of ability to infer the identity in the table. Although quasi identifier attributes are an input of any algorithm to anonymize data, but the formal selection of them is still not researched.

3.1 Selective Quasi identifier attributes Algorithm Steps

Step 1: Nominate set of attributes with full of dependence of person that may be found in different recourses of data (data owners).

Journal of Theoretical and Applied Information Technology
31st July 2016. Vol.89. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Step 2: Find P(S) of the set of nominated attribute.

Step 3: For each element/elements in the P(S), generating table to contain all elements with their total distinct values in table.

Step 4: Element of P(S) with maximum number of tuple will be the set of QI attribute, if there is more than one element, select the element with minimum number of attribute.

4. EXPERIMENTAL RESULT

Different subset of dataset are used for experimental with different scenarios, Figure 2 shows the distinct number of tuples for each combination of Census_Income subset where Figure 3 d.n.o. for each combination of adult dataset . Lastly figure 4 shows combination for both Census_income and Adult dataset. Details of experiment as follow:

Algorithm name: Selective algorithm
Data Set: Census -income
Total number of tuple: 199523
Nominated set: (Age, Sex, Mace)
P(S): (Age, Sex, Mace, (Age, Sex), (Age, Mace), (Sex, Mace), (Age, Sex, Mace))

Table 1: Census –Income-Number of Tuples

Element	Number of Tuples
Age	91
Sex	2
Mace	5
Age, Sex	182
Age, Mace	449
Sex, Mace	10
Age, Sex, Mace	881

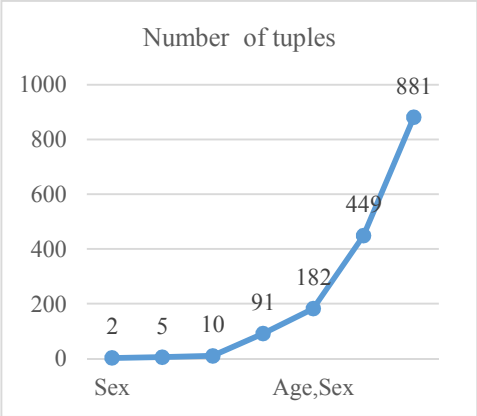

Figure 2: Census -Income Number of Tuples

Data Set: Adult dataset
Total number of tuple: 32561
Nominated set (Zip,Sex,Race)
P(S): (Zip, Sex, Race,(Zip,Sex), (Zip, Race),(Sex, Race), (Zip, Sex, Race))

Table 2: Adult- Number of Tuples

Element1	Number of Tuples
Sex	2
Race	5
Sex, Race	10
Zip	21648
Zip, Sex	22019
Zip, Race	21942
Zip, Sex, Race	22188

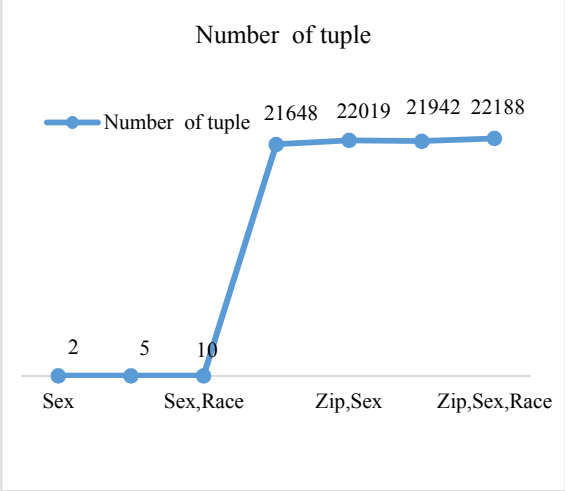

Figure 3: Adult -Number of Tuples

Journal of Theoretical and Applied Information Technology
31st July 2016. Vol.89. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

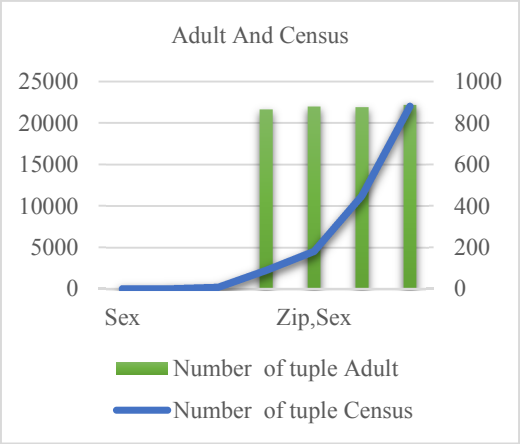

Figure 4: Adult and Census -Income Number of Tuples

From experimental we found that only one attribute with highest ability to infer the identity, normally it is one of continuous attribute and by joining any other attribute with it, will increase this ability, namely, in the adult dataset it is Age attribute, and Zip attribute in Census dataset. To overcome the problem of continuous attribute we proposed decomposed attribute algorithm.

5. DECOMPOSER ALGORITHM FOR DATA REPRESENTATION

Although k-anonymity is a concept that applies to any kind of data, for simplicity its formulation considers data represented by a relational table. Formally, let A be a set of attributes, D be a set of domains, and Dom: A → D be a function that associates with each attribute A ∈ A a domain D=Dom (A) ∈ D, containing the set of values that A can assume. A tuple to over a set $\{A_1. . . A_p\}$ of attributes is a function that associates with each attribute Ai value v ∈ Dom (Ai), i=1. . . P.

DEFINITION 1: (Relational table) let A be a set of attributes, D be a set of domains, and Dom: A → D be a function associating each attribute with its domain. A relational table T over a finite set {A1, . ,A_p}⊆ A, of attributes, denoted T(A1, . . , A_p) is a set of tuples over the set $\{A_1, . , A_p\}$ of attributes. Notation Dom (A, T) denotes the domain of attribute A in T, |T| denotes the number of tuples in T, and t represents the value v associated with attribute A in T. Similarly, t denotes the sub-tuple of t containing the values of attributes $\{A_1. . . A_k\}$. By extending this notation, T represents the sub-tuples of T containing the values of attributes $\{A_1 . . . A_k\}$, that is the projection of T over {A1 . . . A_k}, keeping duplicates.

DEFINITION 2: (Domain generalization relationship) let Dom be a set of ground and generalized domains. A domain generalization relationship, denoted ≤ D, is a partial order relation on Dom that satisfies the following conditions:

C1: ∀Di, Dj, Dz ∈ Dom: Di ≤ D Dj,Di ≤ D Dz ⇒ Dj ≤ D Dz ∨ Dz ≤ D Dj

C2: all maximal elements of Dom are singleton

Condition C1 states that for each domain Di, the set of its generalized domains is totally ordered and each Di has at most one direct generalized domain, Dj. This condition ensures determinism in the generalization process. Condition C2 ensures that all values in each domain can always be generalized to a single value. The definition of the domain generalization relationship implies the existence, for each domain D ∈ Dom, of a totally ordered hierarchy, called domain generalization hierarchy and denoted DGHD. Each DGHD can be graphically represented as a chain of vertices, where the top element corresponds to the singleton generalized domain, and the bottom element corresponds to D. Figure 5: shows an example of DGH.

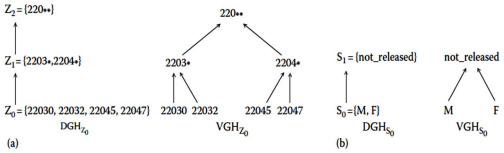

Figure 5[1]: Examples of DGH

Figure 5 shows an example of domain generalization hierarchies for attributes ZIP, Sex, and Marital Status.

DEFINITION 3: (Value generalization relationship) denoted ≤V, can also be defined that associates with each value Vi ∈Di a unique value Vj ∈Dj, where Dj is the direct generalization of Di. The definition of the value generalization relationship implies the existence, for each domain D ∈ Dom, of a partially ordered hierarchy, called value generalization hierarchy and denoted VGHD. Each VGHD can be graphically represented as a tree, where the root element corresponds to the unique value in the top domain in DGHD, and the leaves correspond to the values in D. Figure 4.2 shows an example of value generalization

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

hierarchies for attributes ZIP, Sex, and Marital Status.

Figure 6 shows an example of value generalization hierarchies for attributes ZIP, Sex, and Marital Status.

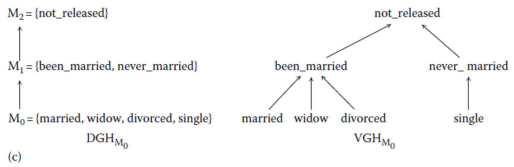

Figure 6[1]: example of value generalization hierarchies

data set sometimes contains compact values as one attribute, like zip code system which includes (State- city-local address) and telephone number system (country code- area code- personal number) those attributes used with other attribute to join the data of an individual so as to infer the Identity of individual, So decompose algorithm work only with a specific type of attribute to split the data into many attribute

5.1 Code Systems for Numbering

It is clear that any international code must have numbering system, for example, Telephone Code Number System – Malaysia (601)

6	0	1	0....0

Another example Postal or zip code System – Indonesia (1240 or 1241

1	2	4	0	0....0

1	2	4	1	0....0

If we split the data in two different table the first one contain the classification according to the system number and the rest of value in another attribute with the same name, for zip code it can be decomposed into 2 digit number as the system in US to represent regional class of zip code, and the rest of code will represent local zip code. Table 4 represent distinct values of sample table 3 for the same number of a tuple.

Table 3: ZIP Code Structure

Zip	State Code	Zip Code
28496	28	496
32214	32	214
32275	32	275
28781	28	781
51618	51	618
51835	51	835
54835	54	835
54352	54	352
59496	59	496
59951	59	951

As a result of decomposition algorithm, Stat Code attribute can substitute by identification number of each state in separated table or state name, new Zip code with less number of digits can be generalized or used in data anonymity

Table 4: Distinct value of table 3 attributes

Zip	State Code	Zip Code
10	5	8

Table 4: show that the ability to identify each tuple is Zip is 100%, but when we split the Zip to state Code and Zip Code the ability is decreased to 50% in state Code and to 80% in new Zip Code.

6. RESULTS AND DISCUSSION

The two algorithms were tested on datasets obtained from UCI Machine Learning Repository: the adult dataset has 32,561 records and 15 attributes of which three attributes (Zip, Race and, Sex) were considered to be quasi-identifiers. The goal is to go to 0 outliers because it contains the value which needs to suppress or change. Total number of distinct record of QI Adult dataset are 16080, total number of adult dataset record 3256. The result of algorithm demonstrated in Figure 7 which shows the relation between outlier before and after applying the algorithm for 3 QID and Figure 8 for 2 QID and Figure 9 appalling for only zip cod.

Journal of Theoretical and Applied Information Technology
31st July 2016. Vol.89. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

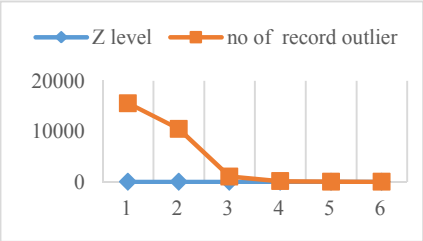

Figure 7: Result of Outlier for Zip, Sex, Rase QI= 3

Result after method of quasi-identifier representation

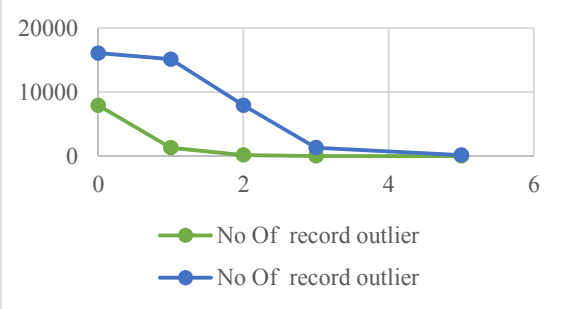

Figure 8: Result of Outlier for sex, Rase QI =2, for 32561 Tuples

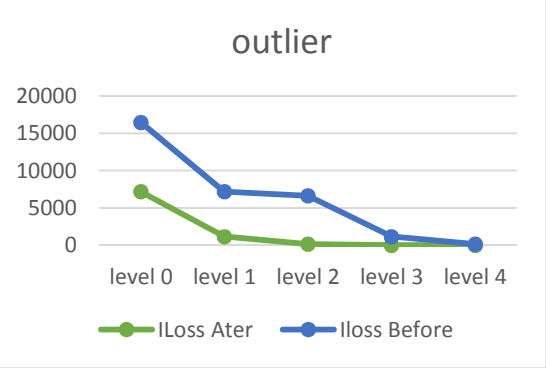

Figure 9: outlier of attribute Zip code before and after appalling the method.

7. CONCLUSION

In this paper, we present simple algorithms for selecting QID [3] followed by decompose algorithm. From the results we show that our method is decreasing loss of information which affect directly of data utility, nonetheless, the minimal set of QI does not imply the most appropriate privacy protection setting because the method does not consider what attributes the adversary could potentially have. If the adversary can obtain a bit more information about the target victim beyond the minimal set, then he may be able to conduct a successful linking attack. So the choice of QI remains an open issue.

REFERENCES:

[1]. Sweeney, L., *k-anonymity: A model for protecting privacy*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002. **10**(05): p. 557-570.

[2]. Fung, B., et al., *Privacy-preserving data publishing: A survey of recent developments*. ACM Computing Surveys (CSUR), 2010. **42**(4): p. 14.

[3]. Motwani, R. and Y. Xu. *Efficient algorithms for masking and finding quasi-identifiers*. in *Proceedings of the Conference on Very Large Data Bases (VLDB)*. 2007.

[4]. Bettini, C., X.S. Wang, and S. Jajodia, *The role of quasi-identifiers in k-anonymity revisited*. arXiv preprint cs/0611035, 2006.

[5]. Gokila, S. and P. Venkateswari, *ASurvey ON PRIVACY PRESERVING DATA PUBLISHING*. International Journal on Cybernetics & Informatics (IJCI) Vol. 3, No. 1, February 2014, 2014.

[6]. Meyerson, A. and R. Williams. *On the complexity of optimal k-anonymity*. in *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2004. ACM.

[7]. Machanavajjhala, A., et al., *l-diversity: Privacy beyond k-anonymity*. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. **1**(1): p. 3.

[8]. Bayardo, R.J. and R. Agrawal. *Data privacy through optimal k-anonymization*. in *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*. 2005. IEEE.

[9]. Li, N., T. Li, and S. Venkatasubramanian. *t-closeness: Privacy beyond k-anonymity and l-diversity*. in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. 2007. IEEE.

[10]. Sun, X., et al., *Enhanced p-sensitive k-anonymity models for privacy preserving data publishing*. Transactions on Data Privacy, 2008. **1**(2): p. 53-66.

[11]. i, T. and N. Li, *Towards optimal k-anonymization*. Data & Knowledge Engineering, 2008. **65**(1): p. 22-39.