International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:6, 2007

Balanced k-Anonymization

Sabah S. Al-Fedaghi

***Abstract*—**The technique of k-anonymization has been proposed to obfuscate private data through associating it with at least k identities. This paper investigates the basic tabular structures that underline the notion of k-anonymization using cell suppression. These structures are studied under idealized conditions to identify the essential features of the k-anonymization notion. We optimize data k-anonymization through requiring a minimum number of anonymized values that are balanced over all columns and rows. We study the relationship between the sizes of the anonymized tables, the value k, and the number of attributes. This study has a theoretical value through contributing to develop a mathematical foundation of the k-anonymization concept. Its practical significance is still to be investigated.

***Keywords*—** Balanced tables, k-anonymization, private data.

I. Introduction

TO protect individual privacy, personal identifying information must be appropriately processed before releasing data. Personal identifying information is private information that links a record to an identified person. It is commonly known, that de-identifying the data does not provide a satisfactory mechanism to guarantee the anonymity of the released private information [14] [8]. De-identifying data refers to stripping it of personal identifying information.

In general, many privacy-protecting methods can be utilized in this context including randomization, cryptography, and anonymity. Information anonymity can be classified into two categories: private information anonymity and non-private information anonymity [1]. We are concerned here with private information anonymity; that is, the anonymity of information that refers to an identifiable individual (person).

The k-anonymization method involves restricting the release of information in a relational table to ensure data privacy while preserving the integrity of the released data. The problem is how to disclose personal identifying data, while preventing identity disclosure [1]. The notion of identity disclosure originated in the area of statistical databases [5].

The columns of a relational table can be categorized into three types: columns that explicitly identify individuals, columns containing potentially identifying information that could be linked with other data sets to re-identify (quasi-identifying columns) and columns containing no identifying information [13]. Quasi-identifying columns are these attributes that in combination can be used to identify an individual. In the k-anonymization technique quasi-identifying columns are transformed through such techniques as generalization and suppression in such a way that every record is indistinguishable from at least k-1 other records.

Generalization is an aggregation of information that is performed through making the data less precise utilizing a taxonomy tree for categorical data or discretization of continuous, numerical values. It involves changing specific values into less specific but semantically consistent values (e.g., "birth date" can be generalized to "birth year"). Suppression is an anonymization method where information is deleted. Our concern here is the type of suppression that results in deleting cell values of tables, in contrast to deleting rows or columns. This type of suppression "consists of omitting (or replacing by an asterisk) the necessary cells to guarantee that an external attacker cannot deduce the sensitive information." [6].

Al-Fedaghi is an associate professor in the Computer Engineering Department, Kuwait University, Kuwait (phone: 965-952-045; fax::965-483-9461, e-mail: sabah@eng.kuniv.edu.kw).

II. Related Works

There are several k-anonymization algorithms proposed in the literature [3] [12] [10] [5]. Here we review a non-exhaustive sample of works in this area. The Hundpool and Willenborg algorithm [11] applies generalization and suppression to all 2 and 3-value combinations of attribute values. The datafly approach [15] generates frequency lists and iteratively generalizes those combinations with fewer than k occurrences. Samarati [9] proposed an algorithm to identify all "k-minimal" generalizations, among which reside the optimal k-anonymizations according to certain preference criteria. The Bayardo and Agrawal algorithm [4] starts with a fully generalized data and systematically specializes the dataset into one that is minimally anonymous.

It is known that an optimal anonymization is NP-hard [7]. We note that the notion optimality through a "minimality" of changes (generalization or suppression is not applied more than is necessary), is a very well known concept in the k-anonymization field. In [11], Samarati introduced an algorithm for finding a single minimal k-anonymous full-domain generalization. Bayardo and Agrawal [5] describe an optimal anonymization, as one which minimizes changes in the input data according to a given cost metric. Generally, the generalization methods work iteratively on the taxonomical hierarchy per attribute in order to achieve minimal generalization up through the hierarchy.

III. The Problem and our Contribution

Most of the different proposed cost metrics for the algorithms in the previous section aim at minimizing the amount of information loss resulting from the generalization and suppression operations. We will concentrate in this paper

on the cell suppression method. Usually, information loss is defined in terms of the number of suppressed cells. A desirable property in k-anonymization is to suppress the minimum set of cells in order to assure adequate confidentiality. Our contribution in this paper is to extend the optimality criteria by balancing the suppression evenly among the columns and rows of the table. To illustrate this concept, consider the 2-anonymization of Table I with tuples t1-t18. Tables II and Table III show two possible solutions. Since the cost is usually calculated in terms of the number of asterisks [3], therefore,

TABLE I
ORIGINAL TABLE

	A1	A2	A3	State
t1	0	0	0	FL
t2	0	0	0	CA
t3	0	0	1	NY
t4	0	0	1	IL
t5	0	1	0	FL
t6	0	1	0	CA
t7	0	1	1	TX
t8	0	1	1	IL
t9	1	0	0	AZ
t10	1	0	0	OK
t11	1	0	1	NY
t12	1	0	1	OK
t13	1	1	0	AZ
t14	1	1	0	NJ
t15	1	1	1	TX
t16	1	1	1	NJ
t17	1	1	1	MI
t18	1	1	1	MI

TABLE II
UNBALANCED TABLE

A1	A2	A3	State
0	0	0	*
0	0	0	*
0	0	1	*
0	0	1	*
0	1	0	*
0	1	0	*
0	1	1	*
0	1	1	*
1	0	0	*
1	0	0	*
1	0	1	*
1	0	1	*
1	1	0	*
1	1	0	*
1	1	1	*
1	1	1	*
1	1	1	MI
1	1	1	MI

TABLE III
BALANCED TABLE

A1	A2	A3	State
0	0	0	*
0	0	0	*
•	0	1	NY
0	•	1	IL
0	1	0	•
0	1	0	•
*	1	1	TX
0	•	1	IL
1	*	0	AZ
1	0	*	OK
•	0	1	NY
1	0	*	OK
1	*	0	AZ
1	1	•	NJ
*	1	1	TX
1	1	•	NJ
1	1	1	MI
1	1	1	MI

both solutions of tables 2 and 3 have equal costs. We claim that, the more distributed the suppressed cells in the table, the less there is a loss of information. In the 2-anonymization of Table 2 most information about 'States' is lost; whereas the loss of information is distributed by equal "amount" in Table 3. In this paper, we study some characteristics of balanced k-anonymized tables of the type shown in table 3. In the k-anonymization literature, the extreme cases of balanced/unbalanced tables 2 and 3 rarely appear. Usually, the given examples are in-between structures. Nevertheless, there is no explicit mentioning of the notion of balanced suppression.

The 'balancing of suppression' criteria can also be applied for the rows (tuples). Consider the 2-anonymization of table 4. Tables V and Table VI show two possible 2-anonymization solutions.

TABLE IV
ORIGINAL TABLE

	A1	A2	A3	State
t1	1	2	3	TX
t2	1	2	3	AZ
t3	8	9	3	TX
t4	1	2	7	TX

TABLE V
UNBALANCED TABLE

A1	A2	A3	State
1	2	3	*
1	2	3	*
*	*	*	TX
*	*	*	TX

TABLE VI
BALANCED TABLE

A1	A2	A3	State
*	*	3	TX
1	2	*	*
*	*	3	TX
1	2	*	*

Both solutions have equal costs in terms of the number of suppressed cells. However, the loss of information is not evenly distributed among tuples in table 5, whereas, in table 6 two cells are suppressed in each tuple. We claim that the more distributed the suppressed cells among rows, the less there is a loss of information. Admittedly, the benefit of distributing *'s among tuples is not as obvious as the benefit of distributing *'s among attributes. We can claim that the tuples represent individuals, hence, balancing the distribution of *'s among these individuals put an upper bound on the loss of information of an individual in the anonymized table. In the unbalanced table 5, we lose 75% of the information of individuals t3 and t4, assuming equal "information values" of all cells. In the balanced table 4, there is only a 50% maximum loss of information of any individual. We can construct more elaborate examples in which the difference in information loss is substantially larger.

This paper brig forth this issue of optimality of k-anonymization. We study some characteristics of fully balanced tables where the number of *'s is equal over columns and rows. We analyze three variables: k (of k-anonymization), m (the number of attributes), and the number of tuples. For given k and m, there may not exist such a balanced k-anonymized table.

IV. Balanced Structures

Without loss of generality, we assume that there are no other attributes except the attributes $A = (A_1, \ldots, A_m)$ that are utilized in the process of anonymizing a table. This is typically assumed in most works on k-anunymization. Also, we modify the definition of k-anonymization by assuming that all tuples are distinct. If there are tuples over A that have identical values, then this is clearly an 'easier' case to anonymize in order to achieve k indistinguishable tuples. Also, without loss of generality, we assume that there is no subset of the attributes A that identify tuples uniquely. Otherwise, this subset can be taken as the set of attributes used in our k-anonymization.

Definition: The table $B(A1,\ldots,A_m)$ with a set of attributes , $A= (A_1, \ldots, A_m)$ and n *distinct* tuples, is said to satisfy the balanced *k*-anonymity property if and only if we minimize the number of suppressed cells, such that:

(a) For each tuple t, there are k − 1 other tuples that are indistinguishable from t.

(b) All columns have an equal number of suppressed cells.

(c) All rows have an equal number of suppressed cells.

Thus, we extend the optimality criteria by spreading the suppression evenly among the columns and rows of the table.

Example 1: The following table taken from [3] is a column-balanced table.

TABLE VII
A COLUMN-BALANCED TABLE

Age	Race	Gender	Zip Code
*	White	*	*
*	White	*	*
27	*	Female	92010
27	*	Female	92010

This type of table has the advantages of distributing the amount of loss of information evenly among attributes. However, the *'s are not distributed evenly among tuples. Of course, in actual situations we may not always be able to achieve such structures. Nevertheless, studying the balanced tables is a necessary step in understanding different types of k-anonymized tables.

We will refer to a balanced k-anonymized table B with m attributed as B(k, m).

Theorem: Let B(k, m) be a balanced k-anonymized table, then:

$$n = i*k*m$$

where i is a positive integer, n is the number of tuples, and m is the number of attributes.

Proof: Each round of cell suppression in the anonymization process makes k*m tuples indistinguishable from each other. That is, each suppression step takes k tuples and suppresses one value from the same attribute in these tuples. Since there is an equal number of asterisks in each column, we need k*m tuples to accomplish this distribution.

Eliminating the redundant rows of the balanced k-anonymized table creates what we call, the balanced structure, BKA-structure.

Example 2: Assume binary domains. For B(2, 2), the largest n such that $2^m \geq i*k*m$ is 8 tuples where k = 2, m = 2, and i = 2. A sample table and its balanced 2-anonymized table are shown in table 8 and 9. The corresponding BKA-structure has i* k rows as shown in table 10.

TABLE VIII
UNBALANCED TABLE

	Gender	HIV/AIDS
t1	Male	Negative
t2	*Female*	Negative
t3	Male	*Negative*
t4	Male	*Positive*
t5	*Male*	Positive
t6	*Female*	Positive
t7	Female	Negative
t8	Female	Positive

TABLE IX
BALANCED TABLE

Gender	HIV/AIDS
*	Negative
*	Negative
Male	*
Male	*
•	Positive
•	Positive
Female	•
Female	•

TABLE X
BKA-STRUCTURE

*	
	*
•	
	•

No possible B(2, 2) is feasible with more than eight tuples; i.e., n > i*k*m, where i (the number of rounds of suppression) is 2. Notice that each row in the balanced table 9 represents two tuples in the original table 8. The process of 2-anonymization includes two rounds of suppression:

(1) For attribute Gender: suppressing tuples t1 and t2; and for HIV/AIDS: suppressing tuples t3 and t4. Thus, this round takes four tuples.

(2) For attribute Gender: suppressing t5 and t6; and for HIV/AIDS: suppressing t7 and t8. Thus, this round takes four tuples.

To illustrate this fact, we insert two different symbols, * and •, for suppression as shown in the corresponding BKA-structure table 10.

From the basic BKA-structure, table 10, of B(2,2) we may construct many tables - not necessarily with binary domains. For example, the basic BKA-structure of B(2,2) in which values are left blank, may correspond to the structure and the table shown in tables 11 and 12.

TABLE XI
BKA-STRUCTURE

*	1950
California	*
•	1960
Florida	•

TABLE XII
BALANCED TABLE

State	Year
Texas	1950
Illinois	1950
California	1990
California	2000
Wisconsin	1960
New York	1960
Florida	1980
Florida	1970

TABLE XIII
A TABLE WITH TERNARY DOMAINS

t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	t15	t16	t17	t18	t19	t20	t21	t22	t23	t24
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
0	0	0	1	1	1	2	2	2	0	0	0	1	1	1	2	2	2	0	0	1	1	1	2
0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	1	2	0	1	2	0

TABLE XIV
B(2, 3)

0	0	0	0	0	0	0	•	*	*	•	1	1	1	1	1	♣	♥	♣	♥	2	2	2	2
0	0	*	*	1	1	•	2	2	0	0	•	1	1	♣	♣	2	2	0	0	♥	1	1	♥
*	*	2	0	•	•	0	1	2	0	1	2	♣	♣	2	0	1	2	1	2	0	♥	♥	0

Example 3: For the table B(2, 3), the largest n such that $2^m \geq i*k*m$ is 6 tuples. Notice that there may exist several versions of B(2, 3) for the same domain.

Example 4: Assume binary domains. For the table B(3, 3), the largest n such that $2^m \geq i*k*m$ is 9 tuples. It is not possible to produce a table of this type. The number of possible (distinct) rows is $2^3 = 8$, which is less than 9. This example raises the possibility of constructing near-balanced k-anonymized tables.

Example 5: Assume binary domains. For the table B(4, 4), the largest n such that $2m \geq i*k*m$ is 16 tuples.

It can be proven that a balanced table for B(4, 4) does not exist, however, space does not allow for such an elaborate proof.

Example 6: Consider table 13 with ternary domains and three attributes. We print the table in the horizontal position to save space. Thus, the first tuple is (0, 0, 0), the second tuple is (0, 0, 1), the third tuple is (0, 0, 2), etc. In B(2, 3) shown as table 14, we have $3^m \geq i*k*m$ or 24 tuples, since i = 4 is the greatest constant that satisfies the constraint. The balanced table in this case can take four rounds of suppression represented in table 14 by the symbols *, •, ♣, and ♥.

V. Conclusion

The balanced k-anonymization structure is an interesting mathematical structure, which may correspond to many tables that can be k-anonymized in a balanced fashion. In our treatment of the subject, we have concentrated on the characteristics of such tables. Further work would develop algorithms to produce these structures for given values of k, m, and domains with different cardinalities.

How these structures can be related to a given relational table? In practice, achieving balanced tables is not practical because they are very rare. However, it is plausible to focus on building maximally balanced anonymized tables. Therefore, possible research work includes two directions:

1- Incorporating the concept of balanced anonymized tables in known anonymization algorithms.

2- Developing a method that maps a given table to its closest BKA-structure. This approach may be feasible for small tables.

References

[1] S. S. Al-Fedaghi,, "*A systematic approach to anonymity*," Proceedings of 3rd International Workshop on Security in Information Systems WOSIS-2005, Miami, May, 2005.

[2] S. S. Al-Fedaghi., G. Fiedler, and B. Thalheim "*Privacy enhanced information systems*," Proceedings of The 15th European-Japanese Conference on Information Modelling And Knowledge Bases, Tallinn, Estonia, 2005.

[3] G. T. Aggarwal, G., K. Feder, R. Kenthapadi, R. Motwani, D. Panigrahy, D. Thomas, A. Zhu, "*k-anonymity: algorithms and hardness*," 2004, http://dbpubs.stanford.edu:8090/pub/2004-24.

[4] R. J. Bayardo and R. Agrawal, "*Data privacy through optimal k-anonymization*" Proc. of ICDE-2005, 2005.

[5] G. Duncan, and D. Lambert, "*The risk of disclosure for microdata*," Journal of Business & Economic Statistics," 7, 1989, pp. 207-217.

[6] J. S. González., "*Improving cell suppression in statistical disclosure control*," Conference of European Statisticians, Skopje, 14-16 March 2001
http://www.unece.org/stats/documents/2001/03/confidentiality/16.e.pdf

[7] A. Meyerson, and R. Williams, "*On the complexity of optimal k-anonymity*," PODS 2004 June 1416, 2004, Paris, France.

[8] L. Sweeney, "*K-anonymity: a model for protecting privacy*," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[9] P. Samarati, "*Protecting respondents' identities in microdata release*.," IEEE Transactions on Knowledge and Data Engineering, 13(6), November/December 2001.

[10] S. Zhong, Z. Yang, and R. N. Wright, "*Privacy enhancing k-anonymization of customer data*," PODS 2005 June 1315, 2005, Baltimore, Maryland.
http://www.almaden.ibm.com/cs/people/bayardo/ps/icde05.pdf

[11] A. Hundpool and L. Willenborg, "*Mu-argus and tau argus: software for statistical disclosure control*," Third Int'l Seminar on Statistical Confidentiality, 1996.

[12] A. Meyerson and R. Williams, "*On the complexity of optimal k-anonymity*," In Proc. of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, 223-228, 2004.

[13] E. Bertino, C. O. Beng, Y. Yanjiang, and R. H. Deng, "*Privacy and ownership preserving of outsourced medical data*," 2005 International Conference on Data Engineering (ICDE), Tokyo, Japan, http://www-scf.usc.edu/~csci586/paper/icde05.pdf.

[14] L. Sweeney, "*Achieving k-anonymity privacy protection using generalization and suppression*," Int'l Journal on Uncertainty, Fuzziness, and Knowledge-Base Systems 10(5): 571-588, 2002.

[15] L. Sweeney, "*Datafly: a system for providing anonymity in medical data. In Database Security XI: Status and Prospects*," IFIP TC11 WG11.3 11th Int'l Conf. on Database Security, 356-381, 1998.