A Privacy Reinforcement Approach against De-identified Dataset

Ci-Wei Lan
Taiwan Research Collaboratory
IBM
Taipei, Taiwan
lancwleepl@gmail.com

Yi-Hui Chen
Dept. of Applied Informatics and Multimedia
Asia University
Taichung, Taiwan
chenyh@asia.edu.tw

Tyrone Grandison
T.J. Watson Research Center
IBM
New York, USA
tyroneg@us.ibm.com

Angus F.M. Huang
Institute of Information Science
Academia Sinica, Taiwan
Taipei, Taiwan
angushuang@iis.sinica.edu.tw

Jen-Yao Chung
T.J. Watson Research Center
IBM
New York, USA
jychung@us.ibm.com

Li-Feng Tseng
Taiwan Research Collaboratory
IBM
Taipei, Taiwan
lftseng@tw.ibm.com

***Abstract*—Protection of individual privacy has been a key issue for the corresponding data dissemination. Nowadays powerful search utilities increase the re-identification risk by easier information collection as well as validation than before. Despite there usually performs certain de-identified process, attackers may recognize someone from released dataset with which attacker-owned information is matched. In this paper, we propose an approach to mitigate the identity disclosure problem by generating plurals in a given dataset. The approach leverages decision tree to help selection of quasi-identifier and several masking techniques can be employed for privacy reinforcement. In addition to different privacy metrics applicability, the approach can achieve better trade-off between data integrity and privacy protection through flexible data masking.**

Keywords- Privacy; quasi-identifier; data mask; microdata protection

I. Introduction

The accessibility of Information and Communication Technology (ICT) enables data collection and dissemination much easier and faster than before. It is also convenient to have analytical results [1] or raw dataset [2] from governmental agencies, hospital, university and corporation etc. Malicious attackers are able to associate known information of someone with these publicly released data and the person's sensitive secrets may be uncovered consequently. How to prevent such re-identification risk from information disclosure has become a great challenge today [3, 4].

According to inference techniques, there are different hacking models including (1) Prosecutor attack: using unique background information to discover confidential secret, for example, attacker knows someone joining a survey in advance and thus can confine the identity search to a small group. (2) Journalist attack: relying on a collection with considerable attributes so that individuals in published analytical results will be re-identified by exactly matching of attribute values. (3) Marketer attack: rather than disclosing specific individual's privacy, a group of population is desired. Attacker aims at having better recall than precision in terms of marketing purpose. Generally, the privacy of individual is disclosed through comparing known information with publicly released data as illustrated in Fig. 1.

Public information

Name	Gender	Date of Birth	ZIP	Height	Weight
Allen Chang	F	41.08.16	11073	~ 150	~ 55
Bill Wang	F	44.03.28	11046	~ 150	~ 55
Cat Lee	F	42.07.01	11058	~ 155	~ 60
Dong Gua	M	55.05.20	11059	~ 165	~ 80
Emily Tsu	F	58.06.18	11073	~ 170	~ 80
Fanny Lu	F	78.08.13	11059	~ 170	~ 85
Gary Lee	F	51.11.15	11073	~ 155	~ 65
Helen Tsai	F	46.12.10	11058	~ 155	~ 65
Ivan Chen	M	61.11.21	11046	~ 155	~ 70
Jacob Lin	F	45.05.15	11059	~ 145	~ 60
Kim Chen	F	46.12.09	11058	~ 150	~ 50
Linda Lee	F	70.10.20	11046	~ 165	~ 95
Mary Mi	M	63.01.12	11059	~ 160	~ 75

Published analytical results

Gender	Date of Birth	ZIP	BMI	Disease
F	38.04.02	11073	29.1	Diabetes
F	45.12.20	11058	24.7	Hypertension
F	39.10.01	11059	30.8	Diabetes
M	34.12.22	11059	29.7	Diabetes
F	52.12.06	11073	26.1	Stroke
F	44.03.28	11046	25.5	Stroke
F	44.04.26	11058	27.4	Stroke
F	58.04.06	11046	25.6	Hypertension
M	34.03.29	11073	29.5	Stroke
F	52.08.25	11058	28.5	Stroke
F	32.05.20	11073	32.9	Diabetes
F	44.03.15	11046	27.1	Hypertension
M	32.11.12	11059	29.5	Diabetes

Figure 1. An example of privacy disclosure

In addition to the linkage between publicly released data and known information, singular cardinality of matched record is another root cause of people's sensitive information disclosure. For example, there is only one record whose value of tuple (Gender, Date of Birth, ZIP, BMI) is correspondent with Bill Wang's in Figure 1. The elements in the tuple are called quasi-identifiers which refer to the intersection set of attributes in publicly released data and known information. There are several privacy evaluation metrics defined upon quasi-identifier to address singular cardinality issue such as k-anonymity [5], l-diversity [6] and t-closeness [7] etc. The metrics provide a quantitative measurement by computing minimal pluralities of quasi-identifier's value combinations in a dataset. On the other hand, a lot of micro-data protection techniques, e.g. masking and synthetic data generations [8], are also available for dataset transformation. However it remains difficult to make a given dataset fulfilling a specified privacy protection metric with reasonable efficiency and appropriate trade-off as well. Fig. 2 exemplifies the dilemma of data refinement towards higher confidence of securing individual secrets. Obviously refined dataset is with strong privacy protection meanwhile nearly useless information disclosure.

Original dataset

Gender	Age	ZIP	BMI	Disease
F	62	11073	29.1	Diabetes
F	55	11058	24.7	Hypertension
M	61	11059	30.8	Diabetes
F	66	11059	29.7	Diabetes
M	48	11073	26.1	Stroke
F	56	11046	25.5	Stroke
M	56	11058	27.4	Stroke
F	42	11046	25.6	Hypertension
M	66	11073	29.5	Stroke
M	48	11058	28.5	Stroke
F	68	11073	32.9	Diabetes
M	56	11046	27.1	Hypertension
F	68	11059	29.5	Diabetes

Refined dataset

Gender	Date of Birth	ZIP	BMI	Disease
***	> 60	1****	> 24	Diabetes
***	> 60	1****	> 24	Hypertension
***	> 60	1****	> 24	Diabetes
***	> 60	1****	> 24	Diabetes
***	40 ~ 60	1****	> 24	Stroke
***	> 60	1****	> 24	Stroke
***	> 60	1****	> 24	Stroke
***	40 ~ 60	1****	> 24	Hypertension
***	> 60	1****	> 24	Stroke
***	40 ~ 60	1****	> 24	Stroke
***	> 60	1****	> 24	Diabetes
***	> 60	1****	> 24	Hypertension
***	> 60	1****	> 24	Diabetes

Figure 2. An example of dilemma between data integrity and privacy protection

In this paper, we propose a privacy reinforcement approach to provide the flexibility in terms of trade-off between data integrity and privacy protection. Firstly we leverage decision tree to help clustering of singular records and employ utility function to consider specific masking strategies such as significance of attribute, the most cardinality, the least refinement etc. Next aforementioned micro-data protection techniques are applicable to perform data transformation. Each mate from clustering procedure will go through the process iteratively until meeting the specified metric. The novelty of our approach can be summarized as follows. (1) Proactively privacy protection: Given a specified protection metric and a dataset, the corresponding refinements can be automatically carried out to reinforce the privacy protection. (2) General purpose solution: The approach is not limited to pre-determined data access scenarios. Whoever will use a dataset to do what, the individual privacy therein can be protected without having case-by-case policy configuration. (3) Flexible trade-off: User can take contextual considerations into account with different quasi-identifier selection strategies.

The remaining parts of this paper are organized as follows. Related literature on privacy protection metrics as well as micro-data protection techniques is discussed in Section II. The decision tree based privacy reinforcement approach is presented in Section III and an example is demonstrated in Section IV. Finally, concluding remarks are illustrated in Section V.

II. TYPE STYLE AND FONTS

A. Privacy protection metric

Since de-identification is not sound enough, several metrics have been proposed to enhance privacy protection measurement against linking attacks. Samarati and Sweeney [5] defined the notion of k-anonymity which indicates any given tuple of quasi-identifiers in a dataset will consist of at least k equivalences. Hence, attackers can re-identify an individual from k candidates with the best case, i.e. with 1/k probability of identity disclosure. The higher k-anonymity a dataset is the better resistance to linking attacks the dataset is capable of.

However, Machanavajjhala et al [6] proposed another measurement perspective called l-diversity to consider the variety of sensitive information. Attackers need not to exactly know which record in a dataset maps to the one known in advance. Once sensitive information of all matched records is the same to each other, attackers can conclude that the individual they try to probe must be with the identical secret data. In other words, no matter how large the k-anonymity a dataset can be, there is still vulnerability if it is with one-diversity. Fig. 3 shows an example to illustrate k-anonymity and l-diversity metrics.

Quasi-identifier: Gender, Age, ZIP, BMI; Sensitive attribute: Disease

Gender	Age	ZIP	BMI	Disease	
F	56	11046	24 ~ 28	Hypertension	k=3, l=2
F	56	11046	24 ~ 28	Hypertension	
F	56	11046	24 ~ 28	Stroke	
F	68	11073	> 28	Diabetes	k=4, l=1
F	68	11073	> 28	Diabetes	
F	68	11073	> 28	Diabetes	
F	68	11073	> 28	Diabetes	
M	48	11058	> 28	Stroke	k=1, l=1
M	48	11073	24 ~ 28	Stroke	k=1, l=1
M	56	11058	24 ~ 28	Hypertension	k=2, l=2
M	56	11058	24 ~ 28	Stroke	
M	61	11059	> 28	Diabetes	k=2, l=2
M	61	11059	> 28	Stroke	

The dataset is with 1-anonymity 1-diversity

Figure 3. k-anonymity and l-diversity metrics

Li et al [7] gave the t-closeness metric to further look into the distribution of sensitive information within a set of identified candidates. The rationale comes from the probability of successfully guess of secret data. In healthcare contexts, the distribution of released sensitive information is usually imbalanced, e.g. whether positive or negative to some bio-test. In this case, attackers can have more confidence to believe an individual is with the same secret as the majority. Hence, t-closeness attempts to refine any set of identified candidates so that the difference of sensitive information distribution between the set of identified candidates and the whole dataset is less than a threshold t.

B. *Micro-data protection technique*

In order to refine a given dataset towards better resistance to privacy sniffer, a lot of data scrambling methods have been developed [8]. Generally there are two different strategies according to whether using fictitious data or not. Synthetic data generation techniques rely on putting simulated fakes into original dataset where the key statistical characteristics are preserved as many as possible. The re-identification risk will be proportional to the amount of synthetic data, i.e. the more generated data are the lower probability of real individual's privacy disclosure is. Bootstrap [9] is a fully synthetic method by mimicking probabilistic distribution of attributes from sampling of original dataset. The sampling records will be replaced with mocked ones. Nevertheless bootstrap is applicable to continuous attribute only due to the premise of calculating distribution function. There are more full synthesis approaches available such as Cholesky Decomposition [10], Multiple Imputation [11], Maximum Entropy [12] and Latin Hypercube Sampling [13] etc. Sometimes it is unable to produce a complete row of data and partially twisting a real one is an alternative solution. Federal Committee on Statistical Methodology [14] proposed a blank-and-impute method where original values are replaced with appropriate function outputs such as median or average. Random Response [15] is another similar technique which frames the problem of synthetic data generation as dealing with outlier issue reversely. Knowing how values between a set of attributes will relate to each other, it is able to change those right tuples to another onces which are consistent with the knowledge.

In addition to creating replica, masking is also a useful technique for micro-data protection. Rather than feeding original dataset with generated fakes, different operations on dataset are performed to keep statistical property in masking techniques. Generalization [16] modifies original dataset by hiding details, for example, changing 5 digits zip code 11058 to 4 digits 1105*. The idea tries to enlarge search space of linking individual to some record as well as to expand the size of equivalent tuples. However, there needs to define the generalization hierarchy of attributes in advance and some attributes may have flat abstractions such as gender. Suppression [17] is another solution by removing attribute values directly if they are significantly sensible quasi-identifiers, i.e. attackers can probably infer the identity of individual by knowing the attribute's values. Recoding [18, 19] can be viewed as special generations with threshold judgment. Given a threshold value of attribute, e.g. 180 mmHg for systolic blood pressure, all values which are greater than 180 are refined as > 180. Therefore, the uniqueness of extreme value will be concealed so as to prevent people with special features from being recognized easily. Besides, there are also other perturbative masking techniques like Resampling [20], Rounding [15] and Swapping [21] etc.

LeFevre [22] et al proposed a method named Incognito to perform full-domain k-anonymity. They start the anonymization process from single attribute first and incrementally aggregate qualified attribute hierarchies to potential more attribute anonymization. This bottom-up approach is somewhat better than searching all possible combinatorial attribute hierarchies. However, the full-domain feature, i.e. masking attribute value of all records to the same abstraction layer, will cause more information loss than tuning partially. Bayardo and Agrawal [23] presented another solution by framing the problem of finding out optimal generalization or suppression as searching the power set of all abstraction hierarchies. Similarly, it requires full-domain masking of dataset and may be hard to keep original statistical property.

III. A DECISION TREE BASED PRIVACY REINFORCEMENT APPROACH

In order to strengthen identity secrecy of de-identified dataset, we propose a decision tree based approach to remove singularity phenomenon as illustrated in Fig. 4. If a given dataset doesn't meet the specified metric, the refinement process will execute by merging non-plural records. The process starts by constructing the decision tree with given dataset and mating singular records according to defined utility function. For each mate, selecting the quasi-identifier from decision tree with most benefits and masking the records correspondingly. Once all mates fulfill specified metric, it will return the privacy reinforced dataset. Otherwise, each mate will go through the refinement process individually until stopping criteria is met.

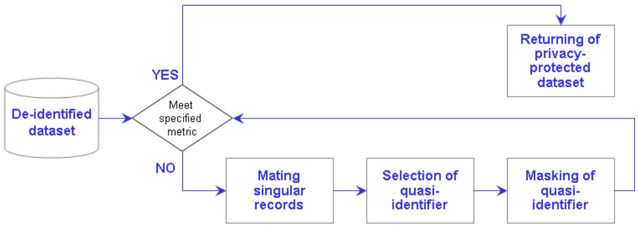


Figure 4. The process of privacy reinforcement approach

Without loss of generality, we can define quasi-identifiers as the intersection elements of attributes in a given dataset and attacker known information. The remaining attributes in the dataset are seen as sensitive information as described in Figure 3. Provided that records with identical values of quasi-identifier are monotonic in terms of certain privacy protection metric, they are called singular records. For example, rows with tuple (Gender, Age, ZIP, BMI) = (F, 68, 11073, >28) in Figure 3 are singular in terms of l-diversity but non-singular in terms of k-anonymity.

Since there are combinatorial possibilities of merging singular records, we leverage decision tree to help dataset clustering and employ utility function to measure gained benefits with reasonable efficiency. The following algorithm presents detailed procedures of how to perform dataset refinements towards specified metric. If there are multiple sensitive information, decision tree construction should be performed individually and utility function needs to take all benefits into consideration as a whole.

De-identified dataset refinement algorithm
Input:
D: a de-identified dataset with defined quasi-identifiers *Q* and sensitive information *S*
P: a reinforcement goal in terms of privacy protection metrics
U: a utility function to measure benefits against merging singular records
Output: Refined de-identified dataset

1: CONSTRUCT a decision tree *T* with *D* as training set, *Q* as classifying attributes and *S* as class label
2: FOREACH singular path *SP* in *T*
3: CALCULATE benefits of merging SP_i and SP_j with *U*, where $i \neq j$
4: ENDFOREACH
5: FOREACH mate $M = (SP_{m1}, SP_{m2}, \ldots SP_{mi})$ with the corresponding subtree T_m
6: SELECT the classifying attribute *C* from T_m, where *C* can differentiate at least a pair of SP_{mi} and SP_{mj} and $U(C)$ gains most benefits
7: Mask *C* of records in T_m and obtain new sub-dataset D_m
8: IF D_m meets *P*
9: RETURN D_m
10: ELSE
11: GOTO 1 with D_m, *P* and *U*
12: ENDFOREACH

In addition to cluster dataset with reasonable efficiency, the adoption of decision tree also guarantees the correctness of quasi-identifier selection, i.e. Step 6 in aforementioned algorithm.

Claim: For any subtree T_m with singular path SP_{m1}, $SP_{m2}, \ldots SP_{mi}$, selecting a classifying attribute which can differentiate at least a pair of SP_{mi} and SP_{mj} can decrease the size of attribute set where their values in SP_{mi} and SP_{mj} are distinct.
Proof:
Without loss of generality, assuming classifying attribute *C* in T_m can differentiate SP_{mi} and SP_{mj}

Let *DA* be the set of attributes whose values in SP_{mi} and SP_{mj} are distinct

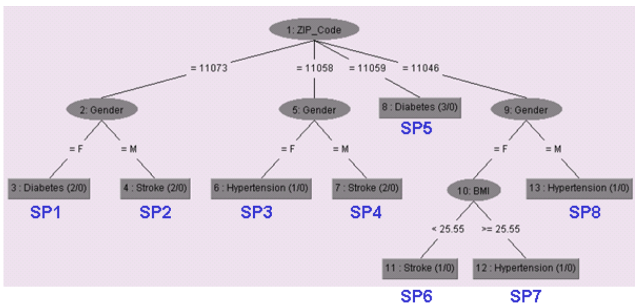
The goal of merging SP_{mi} and SP_{mj} is to make $|DA| = 0$

Obviously *C* is in *DA* so masking *C* in SP_{mi} and SP_{mj} can decrease the size of $|DA|$ by 1

On the other hand, utility function in our design allows more flexibility than other previous work. It is able to consider both cardinal and ordinal utilities at the same time, for example, the importance of attribute, the size of attribute set where their values in distinct singular paths are different etc.

IV. Example and Discussions

In order to better describe the design of our approach, we use the original dataset in Figure 2 as an example. The privacy protection metric is set to 2-diversity and the utility function is defined as 1 / (number of records * distance sum of singular paths). Fig. 5 illustrates the 1st decision tree with original dataset as well as mating decisions.



Singular Path	Most Benefits	Mating Path	Final Mate
SP1	0.25	SP2	M1
SP2	0.25	SP1	
SP3	0.33	SP4	M2
SP4	0.33	SP3	
SP5	0.175	SP3	
SP6	0.5	SP7	M3
SP7	0.5	SP6	
SP8	0.25	SP6	

Figure 5. The first iteration of exemplified singular path mating

For the mate M1, M2 and M3, the value of classifying attribute Gender, Gender and BMI will be masked respectively. By evaluating the corresponding dataset D1, D2 and D3, none of them meet the 2-diversity so they need to go through the process individually. Fig. 6 shows the complete iteration with D1 and Fig. 7 presents the final refinement results.

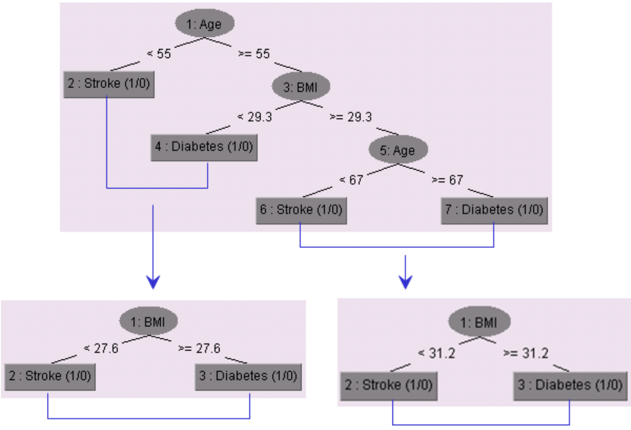


Figure 6. The complete iteration of mate M1

Original dataset

Gender	Age	ZIP	BMI	Disease
F	62	11073	29.1	Diabetes
F	55	11058	24.7	Hypertension
M	61	11059	30.8	Diabetes
F	66	11059	29.7	Diabetes
M	48	11073	26.1	Stroke
F	56	11046	25.5	Stroke
M	56	11058	27.4	Stroke
F	42	11046	25.6	Hypertension
M	66	11073	29.5	Stroke
M	48	11058	28.5	Stroke
F	68	11073	32.9	Diabetes
M	56	11046	27.1	Hypertension
F	68	11059	29.5	Diabetes

Refined dataset

Gender	Age	ZIP	BMI	Disease
F	< 63	11073	> 26	Diabetes
F	< 63	11073	> 26	Stroke
F	> 65	11073	> 29	Stroke
F	> 65	11073	> 29	Diabetes
M	> 55.5	1105*	> 24.6	Diabetes
M	> 55.5	1105*	> 24.6	Diabetes
M	> 55.5	1105*	> 24.6	Stroke
M	> 55.5	1105*	> 24.6	Diabetes
M	< 55.5	11058	> 24.6	Hypertension
M	< 55.5	11058	> 24.6	Stroke
F	> 41	11046	< 27.2	Stroke
F	> 41	11046	< 27.2	Hypertension
F	> 41	11046	< 27.2	Hypertension

Figure 7. The comparison between original and refined datasets

The exemplified demonstration shows the flexibility of dataset refinements and the proposed approach is able to adopt any preference consideration through defining specific utility function. On the other hand, decision tree provides a divide and conquer scheme to deal with anonymization and the refinements can associate with given dataset's distribution rather than fixed masking configurations. In a word, the refined dataset will be closer to the original one while has less privacy concerns.

V. CONCLUSIONS

People are aware of potential risks from distributing personal information to 3rd party organizations. The advances of IT technology cause the situation more dangerous than before. The simple de-identified process by removing recognizable columns from a dataset is not enough for re-identification attacks. In this paper, we proposed a decision tree based privacy reinforcement approach. A dataset can be split into different clusters with corresponding quasi-identifiers and sensitive information. While masking singular records towards specified privacy protection metric, it is able to consider the preference by introducing utility function in mating process. The refinements can be carried out by the divide and conquer scheme with reasonable efficiency. Despite the optimality is not promised, several practical advantages are available including the extensive accommodation of masking techniques as well as privacy protection metrics, the flexibility of masking preference and the dataset-dependent refinement.

REFERENCES

[1] U.S. Census Bureau. (2011). U.S. Census Bureau. http://www.census.gov/.

[2] UCI. (2011). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/

[3] T. C. Rindfleisch. 1997. Privacy, information technology, and health care. Communication of ACM, Vol. 40, Issue 8, 92-100.

[4] L. Brandimarte (2009). Privacy Concerns and Information Disclosure: An Illusion of Control Hypothesis. http://www.heinz.cmu.edu/research/347full.pdf

[5] P. Samarati and L. Sweeney. 1998. Generalizing data to provide anonymity when disclosing information. In Proc. of the 17th ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, 188.

[6] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. 2007. L-diversity: Privacy beyond k-anonymity. ACM Transaction on Knowledge Discovery from Data (TKDD), Vol. 1, Issue 1.

[7] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. 2007. L-diversity: Privacy beyond k-anonymity. ACM Transaction on Knowledge Discovery from Data (TKDD), Vol. 1, Issue 1.

[8] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati (2007). Microdata Protection. Advances in Information Security. Springer.

[9] S.E. Fienberg (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical Report 611, Department of Statistics, Carnegie Mellon University.

[10] J.M. Mateo-Sanz, A. Martinez-Balleste and J. Domingo-Ferrer (2004). Fast generation of accurate synthetic microdata. In Domingo-Ferrer J, Torra V, editors, Privacy in Statistical Databases, vol. 3050 of LNCS, pp. 298-306. Springer, Berlin Heidelberg.

[11] T.E. Raughnathan, J.P. Reiter and D.B. Rubin (2003). Multiple imputatoin for statistical disclosure limitation. Journal of Official Statistics, Vol. 19, No. 1, pp. 1-16.

[12] S. Polettini and L. Franconi (2002). Simulation methods in data protection: an approach based on maximum entropy. In Proc. of the Internatoinal Conference of the Royal Statistical Society, Plymouth.

[13] A. Florian (1992). An efficient sampling scheme: updated latin hypercube sampling. Probabilistic Engineering Mechanics, Vol. 7, No. 2, PP. 123-130

[14] Federal Committee on Statistical Methodology (1994). Statistical policy working paper 22. USA. Report on Statistical Disclosure Limitation Methodology. Menlo Park, California; London; Amsterdam; Don Mills, Ontario; Sydney.

[15] D.E. Denning. (1982). Inference controls. In Cryptography and Data Security, pp. 331-392. Addison-Wesley Publishing Company, Reading, Massachusetts.

[16] P. Samarati. (2001). Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010-1027.

[17] L.H. Cox. (1980). Suppression methodology and statistical disclosure analysis. Journal of the American Statistical Association, Vol. 75, No. 370, PP. 377-385.

[18] J. Domingo-Ferrer and V. Torra. (2001). A quantitative comparison of disclosure control methods for microdata. In Doyle P, Lane JI, Theeuwes J, and Zayatz L, editors, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam.

[19] J. Domingo-Ferrer J and V. Torra. (2002). Distance-based and probabilistic record linkage for re-identication of records with categorical variables. Butlleti del'Associacio Catalana d'Intelligencia Articial, 27.

[20] J. Domingo-Ferrer, J.M. Mateo-Sanz. (1999). On resampling for statistical confidentiality in contingency tables. Computers & Mathematics with Applications, Vol. 38 No. 11, PP. 13-32.

[21] T. Dalenius and S.P. Reiss SP. (1978). Data-swapping: a technique for disclosure control (extended abstract). In Proc. of the ASA Section on Survey Research Methods, pp. 191-194, Washington DC.

[22] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. (2005). Incognito: efficient full-domain K-anonymity. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data (SIGMOD '05). ACM, New York, NY, USA, pp. 49-60.

[23] R.J. Bayardo and R. Agrawal. (2005). Data Privacy through Optimal k-Anonymization. In Proceedings of the 21st International Conference on Data Engineering (ICDE '05). IEEE Computer Society, Washington, DC, USA, pp. 217-228.