ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Generalization Technique for Data Anonymization in Securing PHR

M.Shanthini[1], R.Karthik [2]

III M.E., Dept of CSE, Karpagam University, Coimbatore, India[1]

Assistant Professor, Dept of CSE, Karpagam University, Coimbatore, India[2]

ABSTRACT*:* Several anonymization techniques, such as generalization and bucketization, have been designed for privacy preserving micro data publishing. Maximizing data usage and minimizing privacy risk are two conflicting goals. Organizations always apply a set of transformations on their data before releasing it. While determining the best set of transformations has been the focus of extensive work in the database community, most of this work suffered from one or both of the following major problems: scalability and privacy guarantee. Differential Privacy provides a theoretical formulation for privacy that ensures that the system essentially behaves the same way regardless of whether any individual is included in the database. In this paper, both scalability and privacy risk of data anonymization are addressed. A scalable algorithm is proposed that meets differential privacy when applying a specific random sampling.

KEYWORDS: anonymization , data mining , itemsets

I. Introduction

The data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise. Data mining works to analyze data stored in data warehouses that are used to store that data that is being analyzed. That particular data may come from all parts of business, from the production to the management. Managers also use data mining to decide upon marketing strategies for their product. They can use data to compare and contrast among competitors. Data mining interprets its data into real time analysis that can be used to increase sales, promote new product, or delete product that is not value-added to the company. Data mining offers great potential benefits for GIS-based applied decision-making. Recently, the task of integrating these two technologies has become of critical importance, especially as various public and private sector organizations possessing huge databases with thematic and geographically referenced data begin to realize the huge potential of the information contained therein.

II. Related work

In [1] authors view the k-anonymization problem from the perspective of inference attacks over all possible combinations of attributes. It shows that when the data contains a large number of attributes which may be considered quasi-identifiers, it becomes difficult to anonymize the data without an unacceptably high amount of information loss.

This is because an exponential number of combinations of dimensions can be used to make precise inference attacks, even when individual attributes are partially specified within a range. Authors provide an analysis of the effect of dimensionality on k-anonymity methods. It concludes that when a data set contains a large number of attributes which are open to inference attacks, we are faced with a choice of either completely suppressing most of the data or losing the desired level of anonymity. Thus, this paper shows that the curse of high dimensionality also applies to the problem of privacy preserving data mining. [2] Author presents an efficient algorithm that generates all significant association rules between items in the database. The algorithm incorporates buffer management and novel estimation and pruning techniques. Author also presents results of applying this algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm. It uses pruning techniques to avoid measuring certain itemsets, while guaranteeing completeness. These are the itemsets that the algorithm can prove will not turn out to be large. There is two such pruning techniques. The first one, called the remaining tuple optimization", uses the current scan position and some counts to prune itemsets as soon as they are generated. This technique also establishes, while a pass is in progress, that some of the itemsets being measured will eventually turn out to be large and prunes them out. The other technique, called the pruning function optimization", synthesizes pruning functions in a pass to use them in the next pass.
In [3] authors proposes and evaluates an optimization algorithm for the powerful de-identification procedure known as -anonymization. A -anonymized dataset has the property that each record is indistinguishable from at least others. Even simple restrictions of optimized -anonymity are NP-hard, leading to significant computational challenges. Authors present a new approach to exploring the space of possible anonymizations that tames the combinatorics of the problem, and develop data management strategies to reduce reliance on expensive operations such as sorting. Through experiments on real census data, it shows the resulting algorithm can find optimal -anonymizations under two representative cost measures and a wide range of. It also shows that the algorithm can produce good anonymizations in circumstances where the input data or input parameters preclude finding an optimal solution in reasonable time. Finally, this uses the algorithm to explore the effects of different coding approaches and problem variations on anonymization quality and performance.

In [4] authors initiate a systematic study of algorithms for discrete optimization problems in the frame-work of differential privacy . Authors show that many such problems indeed have good approximation algorithms that preserve differential privacy; this is even in cases where it is impossible to preserve cryptographic definitions of privacy while computing any non-trivial approximation to even the value of an optimal solution, let alone the entire solution. [5] Authors explore preserving the anonymity by the use of generalizations and suppressions on the potentially identifying portions of the data. Satisfying privacy constraints is considered in conjunction with the usage for the data being disseminated. This allows us to optimize the process of preserving privacy for the specified usage. In particular, it investigates the privacy transformation in the context of data mining applications like building classification and regression models. This work improves on previous approaches by allowing more flexible generalizations for the data. Lastly, this is combined with a more thorough exploration of the solution space using the genetic algorithm framework. These extensions allow transforming the data so that they are more useful for their intended purpose while satisfying the privacy constraints.

In [6] authors start to bridge this gap. It first analyzes k-anonymization methods and show how they fail to provide sufficient protection against re-identification, which it was designed to protect. Authors then prove that, k-anonymization methods, when done safely, and when preceded with a random sampling step, can satisfy differential privacy with reasonable parameters. This result is the first to link k-anonymity with differential privacy and illustrates that “hiding in a crowd of k” indeed offers privacy guarantees. This naturally leads to future research in designing “safe” and practical k-anonymization methods. It observes that this result gives an alternative approach to output perturbation for satisfying differential privacy: namely, adding a random sampling step in the beginning and pruning results that are too sensitive to changing a single tuple. This approach may be applicable to settings other than micro data publishing. It also shows that adding a random-sampling step can greatly amplify the level of privacy guarantee provided by a differentially-private algorithm.

III. PROPOSED ALGORITHM

A. *Description of the Proposed Algorithm:*

A scalable algorithm is proposed that meets differential privacy when applying a specific random sampling. The contribution of the paper is two-fold: 1) we propose a personalized anonymization technique based on an aggregate formulation and prove that it can be implemented in polynomial time; and 2) we show that combining the proposed aggregate formulation with specific sampling gives an anonymization algorithm that satisfies differential privacy. Our results rely heavily on exploring the super modularity properties of the risk function, which allow us to employ techniques from convex optimization.

It shows the mechanism which initially samples each record with probability $1 - \beta$. Then for each sampled record $a \in D$, it outputs an element from the generalization $a+ \cap Lt(D)$ according to the exponential distribution (5) defined by the utility. Note that the sampling is necessary, or otherwise the outputs on two databases with different sizes will be different with probability

The threshold frequency t is chosen at random; otherwise, differential privacy cannot be guaranteed to hold since an element can be frequent in in D but not in D−a. Clearly, the output of the algorithm satisfies

(C1). (or (C1_) for the threshold version). We show that it satisfies approximately (C2) and (in some cases) (C3) (or) (C3_) for the threshold version).

IV. PSEUDO CODE

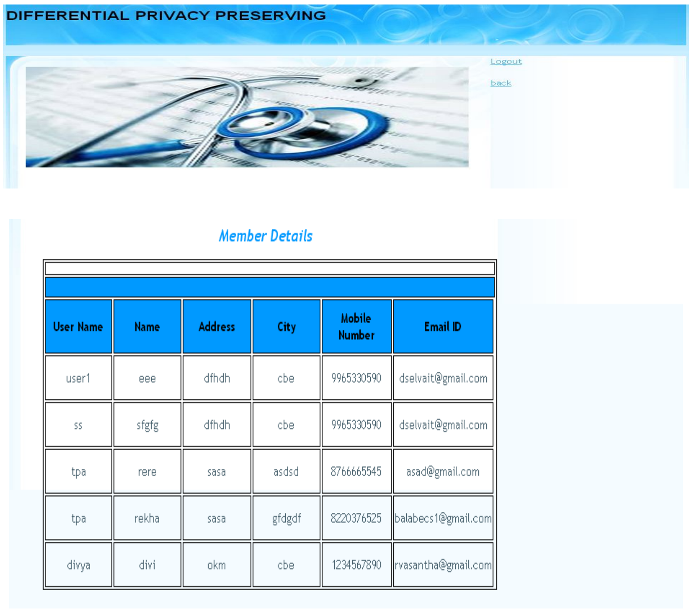
Send probe packet u to neighbors and receive the ack packet.

```
Algorithm 1 A(D, β, θ0)

Input          : a database $D \subseteq L$, a number $\beta \in (0, 1)$, an accuracy , and a constant $\theta 0 \in (0, 1)$
Output  : a subset $G \subseteq L$ satisfying (C1)

Step 1   :  let $q = +\ln \beta \;\; 3\eta(k)(1-\beta)$
Step 2   :  let $t = \theta|D|$ where $\theta$ is chosen randomly in $(\theta 0, 1)$
Step 3   : find the sublattice $Lt(D) \subseteq L$ of t-frequent elements
Step 4   : sample a set $Is \subseteq D$ such that $Pr\,[a \in Is] = 1-\beta$ for all $a \in D$ (independently)  for all $a \in Is$ do
Step 5   : sample $x \in a+ \cap Lt(D)$ with prob. Q _ f a (x); (or sample $x \in a+\cap\{g \in Lta\,(D):u(g) \geq c\}$ with prob.   Q _ ra (x) in   case of the threshold version)
Step 6   : set ga = x
Step 7   : return the (multiset) $\{\perp\} \cup \{ga: a \in Is\}$
Step 8   : End.
```

V. IMPLEMENTATION RESULTS

A differential privacy preserving algorithm is proposed for data disclosure. The algorithm provides personalized transformation on individual data items based on the risk tolerance of the person to whom the data pertains. It first considers the problem of obtaining such a transformation for each record individually without taking the differential privacy constraint into consideration. The goal is to publish an anonymized view of the integrated data such that a data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties. Each provider has additional data knowledge of their own records.

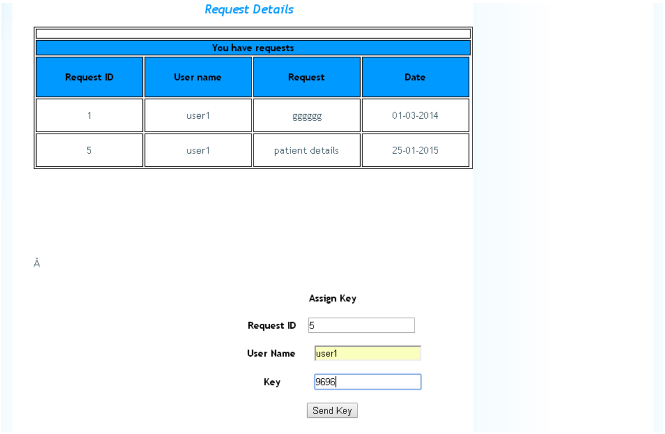
The proposed algorithm is implemented and the resulted output for viewing the group member details is shown in Figure 1 .



Member Details

User Name	Name	Address	City	Mobile Number	Email ID
user1	eee	dfhdh	cbe	9965330590	dselvait@gmail.com
ss	sfgfg	dfhdh	cbe	9965330590	dselvait@gmail.com
tpa	rere	sasa	asdsd	8766665545	asad@gmail.com
tpa	rekha	sasa	gfdgdf	8220376525	balabecs1@gmail.com
divya	divi	okm	cbe	1234567890	rvasantha@gmail.com

Figure 1 : Viewing Group member details



Request Details

You have requests

Request ID	User name	Request	Date
1	user1	gggggg	01-03-2014
5	user1	patient details	25-01-2015

Â

Assign Key

Request ID 5

User Name user1

Key 9696

Figure 2 : Assigning key to each request

Based on the proposed algorithm , a key is assigned to each request which is shown in Figure 2.

VI. Conclusion and Future Work

In this paper both scalability and privacy risk is addressed when identifying the optimal set of transformations which, when carried out on a given table, generate a resulting table that satisfies a set of optimality constraints. Since the problem is NP-hard, we suggested several methods to deal this hardness by utilizing the super modularity properties of the risk function. In particular, an approximation algorithm that computes a nearly optimal solution when the risk threshold is low enough. A scalable Algorithm is proposed that meets differential privacy (with acceptable probability) by applying a specific random sampling. The limitation of the exponential mechanism with respect to the theoretically proved bound on the expected utility . In future we can modify the mechanism such that better utility bounds can be obtained.

REFERENCES

1. C. C. Aggarwal, “On k-anonymity and the curse of dimensionality,” IOSR Journal of Engineering., volume 2,issue no. 3, pp. 337–352, Jun. 2011.
2. R. Agrawal, T. Imieli´nski, and A. Swami, “Mining association rules between sets of items in large databases,” International Journal of Scientific & Engineering Research, volume 58, issue no. 6, pp. 3041–3052, Jul. 2009.
3. D. Applegate and R. Kannan, “Sampling and integration of near log-concave functions,” International Journal of Emerging Technology and Advanced Engineering , ISSN 2250-2459, volume 2, Issue 11, November 2012
4. R. J. Bayardo and R. Agrawal, “Data privacy through optimal kanonymization,” in Third International Symposium on Computer Science and Computational Technology, Jiaozuo, P. R. China, 14-15,August 2010, pp. 151-154.
5. J. Cao, B. Carminati, E. Ferrari, and K.-L. Tan, “CASTLE: Continuously anonymizing data streams,” IEEE Transactions on Secure Computing., volume 8,issue no. 3, pp. 337–352, May 2009.
6. J. Cao, P. Karras, P. Kalnis, and K.-L. Tan, “SABRE: A sensitive attribute bucketization and redistribution framework for t-closeness,” J. International Journal of Engineering, volume. 20, issue no. 1, pp. 59–81, feb 2011.
7. C. Dwork, “Differential privacy,” International Journal of Engineering Research and Applications , ISSN:2248-9622 , Volume. 2, Issue 1,Jan-Feb 2012, pp.967-971.
8. A.M. Khattak, K. Latif, S.Y. Lee, “Change management in evolving web ontologies, Knowledge-Based Systems”, ACM Computer Communication , vol. 28 issue no 4, pp. 5-26, 1999
9. A. Patel, N. Schmidt, “Application of structured document parsing to focused web crawling, Computer Standards & Interfaces” , International journal of data mining ,volume. 12, issue no 2, February 2013.