International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Supermodularity Approach for Differential Data Privacy

Padma L. Gaikwad[1], M. M. Neoghare[2]

[1]Computer Engineering Department, SVIT COE, Chincholi, Sinner, Nasik, Maharashtra, India (ME Scholar)

[2]Computer Engineering Department, SVIT COE, Chincholi, Sinner, Nasik, Maharashtra, India (Assistant Professor)

Abstract: ***Now a day the maximizing of data usage and minimizing privacy risk are two conflicting goals. The organization required set of transformation at the time of release data. While determining the best set of transformations has been the focus on the extensive work in the database community, the scalability and privacy are major problems while data transformation. Scalability and privacy risk of data anonymization can be addressed by using differential privacy. Differential privacy provides a theoretical formulation for privacy. A scalable algorithm is use to find the differential privacy when applying specific random sampling. The risk function can be employ through the supermodularity properties such as convex optimization.***

Keywords: Differential privacy, Scalability, privacy, supermodularity, convex optimization

1. Introduction

Data disclosure is more advantageous in an organization for achieving the privacy and security. To achieve such a goal the organizations apply a set of transformation on a data before releasing it. The data can be microdata contain information about person, a business, or an organization. Data Anonymization is a technique that is used to convert clear text into a non-human readable form. Anonymization uses Generalization and Bucketization techniques. The supermodularity-based differential privacy preserving algorithm provides both scalability and privacy risk by using various algorithms. Data Anonymization is a technique that includes hiding the identities that is called k-anonymity technique. K-anonymity provides accurate data released. K-anonymity technique also use for micro data protection. The l-diversity can overcome the weaknesses of k-anonymity. K-anonymity is not always effective in preventing the sensitive attributes of the record. The technique l-diversity maintains the group of sensitive attributes. Characteristics of l-diversity are to treats all values of attribute in a similar way irrespective of distribution in the data. The t-closeness is one of the techniques ensuring the distance between the distribution of sensitive attributes in a class of records and the global distribution. The transformation includes data suppression, data generalization, and data perturbation. Data suppression removes information from the data. Data generalization can add the information in the form of range such as age into ranges. Data perturbation can help to add noise to the data. The risk utility tradeoff is the main issue regarding data transformation. Generalization and suppuration can reduce the granularity of data representation. One another technique for data privacy is the randomization method. In randomization method noise is added into data for masking the attribute values of records. The added noise must be large so that the individual record value cannot be recovered. This type of technique is developed for the aggregate distribution.

2. Related Work

It is the study of risk-utility tradeoff by using different privacy preserving algorithms. Most of the work can be performed by using optimal transformation before the data gets disclosed. Differential privacy preserving algorithm is used for the data disclosure. This algorithm provides transformation on the individual data items. Such a transformation is based on the risk tolerance of the person to whom the data pertains. An approximation algorithm provides the data transformation within constant guarantees of the optimum. Another algorithm used to data transformation is slightly modified from the approximation algorithm is called Polynomial time algorithm. The supermodular function act as a denominator ratio for the risk function. Thus a fractional program can reduced the number of supermodular function and that can be solved by using polynomial time. Differential privacy considers the problem of data transformation on each record in the database. Differential privacy satisfies the risk-utility mechanism, it maximizes the average utility per record. It is not possible to obtain the differential privacy without considering the maximize utility. Differential privacy provides a mathematical model and it can bound the information gain when an individual is added or removed from the data set D. For achieving the scalability and privacy the data generalization model can be performed under the threshold formulation. The threshold formulation can be categorized into the formal model and informal model. Threshold formulation helps to disclose your personal information. It will collect and uses your personal information to operate the data and deliver the services.

M.R. Fouad proposed an efficient algorithm to address the tradeoff between data utility and data privacy. Maximizing data usage and minimizing privacy risk are two conflicting goals. Our proposed algorithm (ARUBA) deals with the microdata on a record-by-record basis and identifies the optimal set of transformations that need to be applied in order to minimize the risk and in the meantime keep the utility above a certain acceptable threshold. We use predefined models for data utility and privacy risk

throughout different stages of the algorithm. This system does not elaborate more on the impact of different risk and utility models on the performance of our algorithm. Estimating the dictionary of the attacker and the required set of transformations based on incremental disclosure of information is also a subject of future research. [2]

3. Proposed System

A supermodularity-based differential privacy preserving algorithm for data anonymization uses different data generalization techniques based on the threshold formulation. An Informal model and Formal model can be used under the threshold formulation.

3.1 The Informal Model

This model shows the relationship between the risk and utility. The (r, u)-plane can distinguish the risk and utility tradeoff.

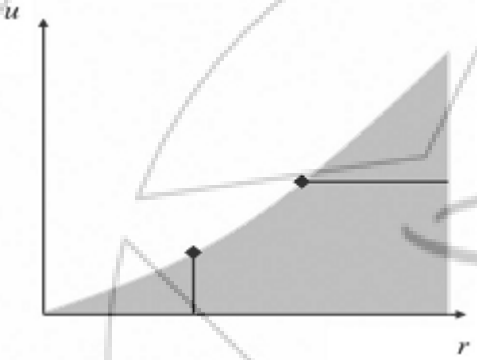

Figure 1: Risk-utility tradeoff

The figure1 shows the shaded region that corresponds to the infeasible points. The vertical line corresponds to all instances whose risk is fixed at a certain level. Similarly the horizontal line corresponds to all instances whose expected utility is fixed at a certain level. The vertical and horizontal line shows the risk-utility tradeoff. Assume that the risk is always below a certain level c.

3.2 Formal Model

This type of model can be work on the basis of Value Generation Hierarchies(VGH's).With the help of VGH we can performed the hierarchical relation.

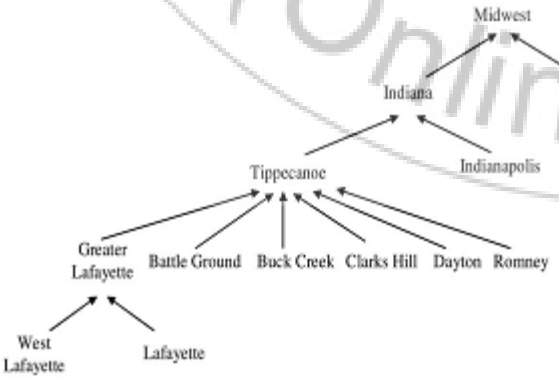

Figure 2: Partial VGH for the city attribute

It provides a utility function as a, u(x) =∑ di (xi) where i=1 to k, k is the number of attributes.

Differential privacy provides a mathematical way to model and bound the information gain when an individual is added to a data set D is a subset of L. Privacy degrades when multiple operations are perfumed on the same set. Differential privacy is advantageous because it degrades privacy in a well controlled manner. Formal model shows the different taxonomies of an attribute. It will generalize the chain product. Formal model shows the two-attribute record in a lattice form. It is formed by chain product by using two attribute. It will show the city and race are the two different attributes. The lattice having three special nodes such as:

1. Feasible nose satisfies the utility constraint,
2. Frontier node has at least one infeasible immediate parent and it is consider as a feasible node.
3. Optimal node is a frontier node that has the minimum risk.

The goal of the lattice representation of formal model is to identify the optimal path. The path is based on only attribute in the record by replacing the attribute value. The figure 3 shows the system architecture of proposed system.

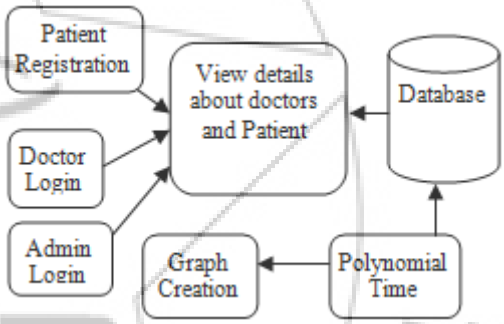

Figure 3: System Architecture

4. Concept

4.1 Polynomial Time

Supermodularity approach for data anonymization includes the polynomial time for completes a task. The time can be calculated for data anonymization. By using this algorithm the normal people can't see the personal details.

4.2 Convex Optimization

Convex optimization is act as a subfield of optimization techniques. It is used in a wide range of disciplines such as many automatic control system, communication and networks, data analysis. Convex optimization is a straightforward approach as a linear programming. It can perform easier optimization than the other type of optimization. It will apply a set of convex functions over a convex set. In data privacy whenever the risk threshold is small, then the convex optimization is used in an approximation algorithm. It approximately maximizes the utility of data within a constant factor from the risk threshold function. A polynomial time algorithm is slightly modified than the approximation algoritm.Polynomial time algorithm produces a reduced number of supermodualar functions that

will maximize the data utility. It is not possible to obtain differential privacy without sacrificing utility maximization. The risk function exhibits certain submodularity properties. The very desirable property of submodular (respectively, supermodular) functions is that they can be minimized (respectively, maximized) in polynomial time [1].

5. Conclusion

A supermodularity-based approach for the data privacy can addressed both the scalability and privacy risk. The set of transformation can apply on the data for maintain the privacy. For achieving the scalability and privacy the proposed system use the risk-utility tradeoff by using optimal set of transformations. The system gave an approximation algorithm for the computation of optimal solution at the time of risk threshold is minimum. By using threshold formulation there are different models introduces the relationship in between the risk and utility. Differential privacy can shows the mathematical model for achieving maximum utility and minimizing privacy risk. Hence it is more popular in database community. The ARUBA and SABRI technology can adopts the information loss measures. In future study we reduce the problem of NP-Hardness.

References

[1] Mohamed R. Fouad, Khaled Elbassioni, "A Supermodularity-Based Differential Privacy Preserving Algorithm for Data Anonymization", *IEEE transaction on Knowledge and Data Engineering July 2014.*
[2] M. R. Fouad, G. Lebanon, and E. Bertino, "ARUBA: A risk-utility based algorithm for data disclosure," in *Proc. VLDB Workshop SDM*, Auckland, New Zealand, 2008, pp. 32–49.
[3] M. R. Fouad, K. Elbassioni, and E. Bertino, "Towards a differentially private data anonymization," Purdue Univ., West Lafayette, IN, USA, Tech. Rep. CERIAS 2012-1, 2012.
[4] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality,"in *Proc. Int. Conf. VLDB*, Trondheim, Norway, 2005,pp. 901–909.
[5] N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu, "Differentially private data release for data mining," in *Proc. 17th ACM SIGKDD*,New York, NY, USA, 2011, pp. 493–501.
[6] G. Lebanon, M. Scannapieco, M. R. Fouad, and E. Bertino,"Beyond k-anonymity: A decision theoretic framework for assessing Privacy risk," in *Privacy in Statistical Databases.* Springer LNCS 4302:217U˝ 232, 2006.
[7] C. Dwork, "Differential privacy," in *Proc. ICALP*, Venice, Italy,2006, pp. 1–12.
[8] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int.Conf. TAMC*, XiŠan, China, 2008, pp. 1–19.
[9] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor,"Our data, ourselves: Privacy via distributed noise generation,"in *Proc. 25th EUROCRYPT*, Berlin, Germany, 2006, pp. 486–503,LNCS 4004.
[10] K. M. Elbassioni, "Algorithms for dualization over products of partially ordered sets," *SIAM J. Discrete Math.*, vol. 23, no. 1,pp. 487–510, 2009.
[11] Frieze, R. Kannan, and N. Polson, "Sampling from log-concave distributions," Ann. Appl. Probab., vol. 4, no. 3, pp. 812–837, 1994.
[12] C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in Proc. IEEE ICDE, Washington, DC, USA, 2005, pp. 205–216.
[13] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in Proc. Int. Conf.VLDB, Vienna, Austria, 2007, pp. 758–769.
[14] G. A. Gratzer, General Lattice Theory, 2nd ed. Basel, Switzerland: Birkhäuser, 2003.
[15] M. Grotschel, L. Lovasz, and A. Schrijver, "Geometric algorithms and combinatorial optimization," in Algorithms and Combinatorics, vol. 2, 2nd ed. Berlin, Germany: Springer, 1993.

Author Profile

Ms. Padma L.Gaikwad has completed her B.E in Computer Engineering from Pune University and currently pursuing Master of Engineering from SVIT Chincholi, Nashik, India

Prof. M. M. Naoghare has completed her B.E in Computer Engineering from College of Engineering, Badnera, Amravati and M.E in Computer Science & Engineering from P.R.M.I.T & R, Badnera, Amravati. She is presently working as an Associate Professor in SVIT Chincholi, Nashik, India