Volume 2, Issue 12, December-2015, pp. 1240-1244 ISSN (O): 2349-7084

International Journal of Computer Engineering In Research

Available online at: www.ijcert.org

An Effective Supermodularity Based Approach for Data Privacy at Storage Level

[1] Annamneni Soujanya,[2] Baburao Kopuri

[1]M.Tech (CSE), Department of Computer Science & Engineering, NRI Institute of Technology

[2] Associate Professor, Department of Computer Science & Engineering, NRI Institute of Technology.

Emails: kbaburao.hodit@gmail.com ,souji.511@gmail.com

Abstract:- In this paper we presented a supermodulaity based approach for data privacy using noval encryption mechanisum in this conncetion the seviority of data privacy at storage level is most considerable so in our presented system lack of data privacy , scalability due to low security algorithm while data transformation so in our proposed system 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.

Keywords: Supermodularity, Differential privacy, Scalability, privacy.

I.INTRODUCTION

In spite of the fact that Data exposure is beneficial for some reasons, for example, research purposes, it might acquire some risk because of security ruptures. Discharging human services information, for instance, however, helpful in enhancing the nature of administration that patients get, raises the odds of personality introduction of the patients. Unveiling the base measure of information (or no information by any means) is convincing particularly when associations attempt to secure the protection of people. To accomplish such an objective, the associations commonly attempt to conceal the character of a person to whom data relates and apply an arrangement of changes to the microdata before discharging it. These changes incorporate (1) data concealment (unveiling the quality ⊥, rather), (2) data speculation (discharging a less particular variety of the first data, for example, in [31]), and (3) data perturbation (adding commotion specifically to the first data values, for example, in [24]). Examining the risk-utility tradeoff has been the center of much research. Determining so as to determine this tradeoff the ideal data change has experienced two noteworthy issues, to be specific, adaptability and protection risk. To the best of our knowledge, the vast majority of the work in deciding the ideal change to be performed on a database before it gets uncovered is wasteful as in expanding the table measurement will considerably compound the execution. In addition, data anonymization strategies don't give enough hypothetical proof that the uncovered table is safe from security ruptures. Anonymization procedures incorporate (1) concealing the characters by making every record vague from in any event k−1 different records [8] (k-anonymity), (2) guaranteeing that the separation between the appropriation of touchy properties in a class of records and the dispersion of them in the entire table is close to t [7] (t-closeness), and (3) guaranteeing that there are at any rate l unmistakable qualities for a given delicate property in each vague gathering of records [26] (l-differences). To be sure, these strategies don't totally anticipate re-distinguishing proof [9]. It appears in [1] that the k-anonymity [8] method experiences the scourge of dimensionality: the level of information misfortune in k-anonymity may not be satisfactory from a data mining perspective on the grounds that the specifics of the between characteristic conduct have an intense uncovering impact in the high dimensional case numerous association works on the ongoing data and they need to individual information for the examination reason. In human services framework, the patient needs to fill all the essential individual information. In the administration part, the individual information incorporates all the important individual data in regards to that individual. Such association can utilize the gathering of the expansive dataset for the optional reason by conceal-

© 2015, IJCERT All Rights Reserved

ing the personalities. To keep up a database protection and give security over the database here the data anonymization system utilized under various suitable instrument and calculations. Since the anonymization technique can just conceal the maybe a couple characters from the table, consequently here the differential security saving instrument help us to give the scientific bound to ensuring the information and once the database bound inside of a reach there are least opportunities to miss the data from the dataset. Before data discharged apply the vital for achieving the privacy and security over the database community. Data disclosure method is more advantageous in an organization for achieving the data privacy and data security. Privacy for the database is becoming a huge problem in many areas such as government, hospitals; many companies etc. Data Anonymization is a one type of technique that is used for conversion of clear text into a non-human readable form. It is used to enable the publication of detail information. Basically data anonymization provides the privacy guarantee for the sensitive data against the various attacks over the database community. To achieve privacy guarantee there are two different techniques such as K-anonymization and ldiversity. K-anonymization is one of the technique which includes the hiding of identities and it is more accurate technology for the data anonymization. There have been no evaluations of the actual re-identification probability of kAnonymized data sets. In k-anonymization each record is distinguishable from k-1 records with respect to certain identifying records. One of the limitations of k-anonymization can overcome the l-diversity. K-anonymization does not provide the privacy guarantees against the attacker using background knowledge. L-diversity is a more powerful technique that can overcome the weaknesses of k-anonymity. K-anonymity is not always effective in preventing the sensitive data of the dataset. The technique of l-diversity is used to maintain the group of sensitive attributes for protecting the data against the background attackers. Characteristics of l-diversity are to treats all values of attribute in a similar way irrespective of distribution in the data. L-diversity is achieved to difficult for sensitive data. It gives different degree of sensitivity. Ldiversity does not consider overall distribution of sensitive values of the record set because of equivalence classes on quasi-identifier. It does not consider semantics of sensitive 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 .In t-closeness the distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database. There are different techniques of an anonymization such as:

1. Data Suppression:-In this technique the information is removed from the data. For example the gender field can be removed from the dataset.

2. Data Generalization:-In this technique the information is coarsened into set or range .For example age of the person can be display in range form.

3. Data Perturbation:-In this technique noise is added directly between the entities. For example the pin code of city can be display in addition of noise form. The differential privacy preserving algorithm provides both scalability and privacy risk by using various polynomial algorithms. Differential privacy provides an interesting and rigorous framework around publishing data. Differential privacy provides to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records. Privacy is important when the contents of a message are at issue and whereas anonymity is important when the identity of the author of a message is at issue. The role of privacy preserving algorithm which prevent the leakage of specific information about person. Sensitive input data is randomized, aggregated, Anonymized and generally contorted to remove any concrete implication about its original form.

II. RELATED WORK

In [12], an algorithm (ARUBA) to address the tradeoff between data utility and data privacy is proposed. The proposed algorithm determines a personalized optimum data transformations based on predefined risk and utility models. However, ARUBA provides no scalability guarantees and lacks the necessary theoretical foundations for privacy risk. A top-down specialization algorithm is developed by Fung et al. [14] that iteratively specializes the data by taking into account both data utility and privacy constraints. A genetic algorithm solution for the same problem is proposed by Iyengar [19]. Both approaches consider classification quality as a metric for data utility. However, to preserve classification quality, they measure privacy as how uniquely an individual can be identified by collapsing every subset of records into one record. The per-record customization nature of our algorithms makes them superior over other algorithms. A personalized generalization technique is proposed by Xiao and Tao [9]. Under such approach users define maximum allowable specialization levels for their different attributes. That is, sensitivity of different attribute values is binary (either released or not released). In contrast, our proposed scheme provides users with the ability to specify sensitivity weights for their attribute values. Perhaps the most related work is the differentially private data release proposed in [28]. In that paper, the authors also consider a product of taxonomies for data generalization, assume some utility function quantifying the information content of the released generalizations, then apply the exponential mechanism to obtain a differentially private mechanism. Their application of the exponential mechanism is done in a somewhat restrictive way in the sense that they do not sample from the space of all generalizations as we

© 2015, IJCERT All Rights Reserved

do. Rather, the sampling is performed in a heuristic way as follows. All the records are put in one group and generalized by the top element (⊥,..., ⊥). Then one of the top elements in the different taxonomies is chosen according to an exponential distribution defined in terms of some utility function. The chosen element is replaced by its children in the corresponding taxonomy. This splits the current group into a number of subgroups, each generalized by an element in the product of the taxonomies. The process is repeated in each of the subgroups. After a predefined number of splits, the count of the number of elements in each of the obtained groups is perturbed by a Laplacian noise. One main restriction of this approach is that the utility function has to be recordindependent. On the contrary, in our formulation we allow the utility function to be different for each record in the database.

III.SYSTEM STUDY

3.1 Proposed System

In this proposed framework the principle concentrate on the issue of discharge factual information around a dataset without trading off the protection of any person. Here the framework can deal with data adaptability and data protection. There are large portions of the procedures accessible that can break the information effortlessly over the database community.The differential security protecting based calculations can give the customized anonymization the offer distinctive protection some assistance with preserving calculation. For data security, an association applies an arrangement of change principles on the database before the utilization of data for the auxiliary reason. The database group contains the delicate data and in addition the quasi-identifier (QI's).The differences and t-closeness can apply the arrangement of standards on the distinctive qualities, for example, touchy data and quasi-identifier independently. The proposed framework essentially works on such a kind of ascribe to accomplish the data protection furthermore expand the data utility.

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

3.3 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. 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

System Architecture

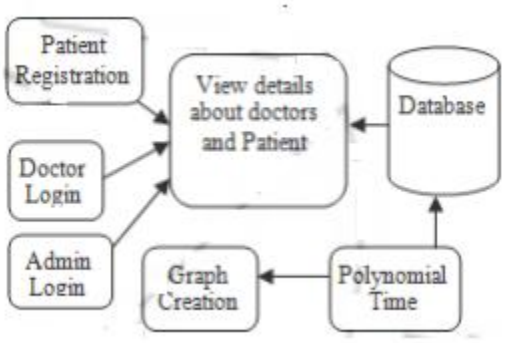

Fig1. System Architecture

IV. DIFFERENTIAL PRIVACY PRESERVING MECHANISM

Differential privacy preserving mechanism aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying or loosing its records. Differential privacy provides privacy preserving algorithm used for data disclosure. Disclosing the minimum amount of information or no information at all is try to protect the privacy of individual to whom data pertains[1].The differential privacy preserving algorithm provide a personalized anonymization on individual data items based on the specific risk tolerance of that data. Differential privacy mechanism can perform the masking operation on individual data, and it allows accurate percentages and trading. An approximation algorithm is deals with hardness under some condition to produce data transformation within constant guarantees of the optimum solution. For achieving differential privacy use the Laplace distribution to add noise probably to add noise in smallest amount required to preserve privacy. f: D Rd K (f, D) =f (D) + [Noise] d The multiplicative factor used in the guarantee of scalable information for higher or lower guarantees of privacy. The noise is depending on the

© 2015, IJCERT All Rights Reserved

factor f and ε, not on the database. Another modified variant of the formulation is a polynomial time algorithm is used for data transformation.

The polynomial time is a one type solvable algorithm and it will refers to time taken required for a computer to solve a problem, where this time is a simple polynomial function of the input. For NPhard problem, there are polynomial algorithms used to solve all problems in NP-algorithm. Polynomial time algorithm can reduce the number of function that will maximize the utility of data. By using polynomial time algorithm, it refers to time taken to complete a task for calculating the time taken for data anonymization. Approximation algorithm work on the smallest value of threshold formulation, over the convex set of optimization. The purpose of approximation algorithm is used for solve linear programming and it is easier optimization than the other algorithm. Threshold value is a minimum or maximum value which serves as a benchmark for comparison or guidance and any breach of which may call for a complete review of the situation or the redesign of a system. Differential privacy provides a mathematical way to model and bound the information gain when an individual is added or removed to or from a dataset D.

It is natural way the privacy degrades when multiple operations are performed on the same set of information and since more information is exposed. But the privacy degrades in a well control manner. A randomized algorithm satisfy the (ε, δ)-differential privacy if, Pr [A (D) ε B] ≤ e ε Pr [A (D′) ε B] + δ For any two data sets D and D′ that differ by at most one record and any subset of outputs B subset Range (A). Differential privacy bound the information gain when an individual is added or removed to or from a dataset. It will give the support for query and requiring that the released data have noise added to ensure that the information for any individual can be sufficiently hidden from the user. It is used for protection purpose Differential privacy ensures for the limited amount of additional risk is incurred by participating in the socially beneficial databases. The removal or addition of any record in the database that does not change the outcome of any analysis by much. That means it ensure the presences of an individual is protected against the attacker's. Differential privacy preserving algorithm work on the basis of sensitivity function. f: D Rd Δf=max||f(x)- f(x′)||1 For all x and x′ differing in at most one element.It captures how great a difference must be hidden by the additive noise. A key technique of randomized rounding of linear relaxations for approximation algorithm is used to rounding a fractional solution x to linear programming relaxation of a problem into an integral solution.

An approximation algorithm maximizes the utility within a constant factor. An approximation algorithm use the Lovasz extension and randomized rounding of a vector extension for finding out the maximum utility. Lovasz extension shows that maximizing a linear function with non-negative coefficients. Convex optimization is one type of techniques which 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 was design for the linear programming. It can perform easier optimization than the other type of optimization. Differential privacy preserving algorithms apply a set of convex functions over a convex set. Convex optimization can be solved globally with similar complexity as linear programming. Many problems can be solved via convex optimization. In data privacy whenever the risk threshold is small, then the convex optimization is used in an approximation algorithm. Threshold value is used for comparison or guidance and any breach of information which may call. It is used for packing integer programs by employing the methods of randomized rounding technique by combining with number of alterations. Steps of Approximation Algorithm:

1. Input: record a, real numbers.
2. Output: Generalization of a.
3. Define lower and upper bound real values for minimum and maximum function
. 4. Execute min () and max () function by using for loop by using till the upper bound.
5. Solve the maximization problem over a convex set. M=max u a (x)
6. Apply randomized rounding extension method over the optimal solution corresponding element a + .
7. Return maximum utility. An Approximation algorithm maximizes the data utility and maintaining risk below certain acceptable threshold value. It can give the guarantees to be close to an optimal solution. It runs in a polynomial time and obtains a good bound on the optimal solution. Randomized rounding method gives an o (log n) approximation.

V. RESULT ANALYSIS

The system is used for the hospital data protection from the attackers. At the time of registration there are two different domains are used for the registration such as personal domain and public domain. Personal domain contains the patient registration and public domain performs the insurance as well as doctor domain. Form the personal domain the system can generate the graph of disease. This graph will be useful for the secondary purpose for investigation of disease.

© 2015, IJCERT All Rights Reserved

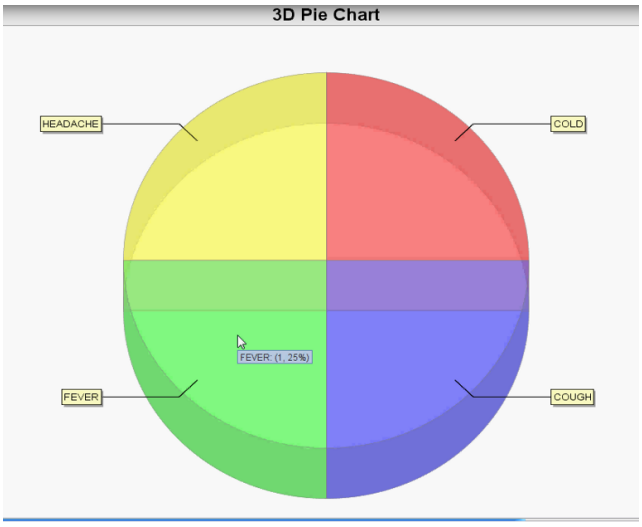

Fig 2. Analysis of dieses

Above graph shows the pictorial view of disease in percentage. This graph of disease can be used for the secondary purpose because it will only display the diseases in percentile ratio not the personal information. Hence the data is anonymized through the differential privacy preserving algorithm and also display such information for the secondary purpose.

VI. CONCLUSION

In this paper we address a supermodularity-based approach for the information privacy can tend to both the scalability and privacy hazard. The arrangement of change can apply the information for keeping up the privacy. For accomplishing the scalability and privacy, the proposed framework utilize the danger utility tradeoff by utilizing the ideal arrangement of changes. The framework gave a guess calculation for the calculation of ideal arrangement at the season of danger limit is least. By utilizing limit detailing, there are diverse models presents the relationship in the middle of the danger and utility. Differential privacy can demonstrate the numerical model for accomplishing most extreme utility and minimizing privacy hazard. Henceforth it is better known in database group.

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, K. Elbassioni, and E. Bertino, "Towards a differentially private data anonymization," Purdue Univ., West Lafayette, IN, USA, Tech. Rep. CERIAS 2012-1, 2012.
[3] N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu, "Differentially prívate data release for data mining," in Proc. 17th ACM SIGKDD,New York,NY, USA, 2011, pp. 493–501.
[4] 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.
[5] K. M. Elbassioni, "Algorithms for dualization over products of partially ordered sets," SIAM J. Discrete Math., vol. 23, no. 1,pp. 487–510, 2009.
[6] C. Dwork, "Differential privacy: A survey of results," in Proc. Int.Conf. TAMC, Xi˘ San, China, 2008, pp. 1–19.
[7] G. Lebanon, M. Scannapieco, M. R. Fouad, and E. Bertino,"Beyond kanonymity:A decision theoretic framework for assessing Privacy risk," in Privacy in Statistical Databases. Springer LNCS 4302:217U 232, 2006.
[8] C. Dwork, "Differential privacy," in Proc. ICALP, Venice, Italy,2006, pp.1–12.
[9] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality,"in Proc. Int. Conf. VLDB, Trondheim, Norway, 2005,pp. 901–909.
[10] 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.
[11] A. Frieze, R. Kannan, and N. Polson, "Sampling from log-concave distributions," Ann. Appl. Probab., vol. 4, no. 3, pp. 812–837, 1994.
[12] B. 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:Birkh¨auser, 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.
[16] J. Cao, P. Karras, P. Kalnis, and K.-L. Tan, "SABRE: A sensitive attribute bucketization and redistribution framework for t-closeness," J. VLDB, vol. 20, no. 1, pp. 59–81, 2011.
[17] 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.

© 2015, IJCERT All Rights Reserved