International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Anonymizing Data Privacy Personalization and the Web

Sandhyarani Hulage[1], K. M. Varpe[2]

[1]Department of Computer Science, RSSOER JSPM NTC, Pune, India

[2]Professor, Department of Computer Science, RSSOER JSPM NTC, Pune, India

Abstract: ***Personal information should not be exposed to other in any case. Personal data of patients is very sensitive. In other hand sharing of health reports are also necessary for research purpose. So here real challenge is how we can share the information in such way that the researchers can get maximum benefits of data without knowing any patients personal information. 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. In this project we used Differential Privacy which 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.***

Keywords: Differential Privacy, security, risk management, data sharing, data utility, anonymity, scalability

1. Introduction

Personal information should not be exposed to other in any case. Personal data of patients is very sensitive. In other hand sharing of health reports are also necessary for research purpose. So here real challenge is how we can share the information in such way that the researchers can get maximum benefits of data without knowing any personal information.

If the privacy is not achieved that system is not secured one, then no one is ready to give their personal details. So if we want to take public survey on people,that system must be secured one. We are going to develop the more secured system. No one can breach the personal data of an individual.

A. Data Anonymization

Data Anonymization is the process of either encrypting or removing personally identifiable information from data sets, so that the people whom the data describe remain anonymous.

Data Anonymization enables the transfer of information across a boundary, such as between two departments within an agency or between two agencies, while reducing the risk of unintended disclosure, and in certain environments in a manner that enables evaluation and analytics pos Anonymization. In the context of medical data, anonymized data refers to data from which the patient cannot be identified by the recipient of the information. The name, address, and full post code must be removed together with any other information which, in conjunction with other data held by or disclosed to the recipient, could identify the patient. De-Anonymization is the reverse process in which anonymous data is cross-referenced with other data sources to re-identify the anonymous data source. Generalization and perturbation are the two popular Anonymization approaches for relational data.

Data masking is one of the most popular approach to live data anonymization. By replacing sensitive data with fake data you will be able to disclose your production data outside of your organization. An ineffective data masking process may result in anonymized but poor quality data, useless for replacing sensitive real data.

The major pitfall of any anonymization process is to focus on masking sensitive data and miss the primary goal: obtain quality data for your test.

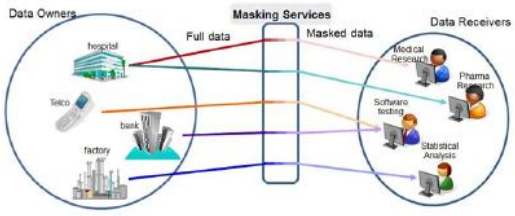


Figure: A scenario of data anonymization Techniques

2. Literature Survey

The study of literature on Data Anonymization: Privacy,Personalization, and the Web in general and in the field of library and information science particular revealed several efforts made by the scholars in different discipline.

The purpose of the literature survey is to collect a lot of number of journal's article about a particular topic like as I have collected many articles on my project topic "Data Anonymization: Privacy, Personalization, and the Web" with abstract. The main aim of this collection is to provide a guideline and brief information of researcher, user and other person who want information about this topic.

Mohamed R. Fouad, Khaled Elbassioni[4] propose a differential privacy preserving algorithm for data disclosure. The algorithm provides personalized transformation on

Licensed Under Creative Commons Attribution CC BY

individual data items based on the risk tolerance of the person to whom the data pertains. We first consider the problem of obtaining such a transformation for each record individually without taking the differential privacy constraint into consideration.

Bin Zhou Yi Han focuses on Continuous Privacy Preserving Pu of Data Streams. Privacy becomes a more and more serious concern in many applications. A large category of privacy attacks is to re-identify individuals by joining the published table with some external tables modeling the background knowledge of users.To battle this type of attacks, the mechanism of k- anonymity was proposed. Adata set is said to be k-anonymous (k 1) if on the quasi-identifier attributes (the minimal set of attributes in the table that can be joined with external information to re-identify individual records), each record is indistinguishable from at least k1other records within the same data set. The larger the value of k, the better the privacy is protected.

Charu C. Aggarwal [1] discusses the effects of the curse of high dimensionality on privacy preserving data mining algorithms. Since k-anonymity models attempt to retain partial information about different dimensions simultaneously they are more open to inference attacks. In many high-dimensional cases, the level of information loss required in order to preserve even 2-anonymity may not be acceptable from a data mining point of view. This is because the specifics of the inter-attribute behavior have a very powerful revealing effect in the high dimensional case. We also conjecture that in such cases, it may be more effective to use perturbation techniques which do not preserve such inter-attribute information but work with aggregate distributions on individual dimensions. Another possibility is to use selective information hiding in conjunction with conceptual reconstruction techniques.

N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu. [11] Focuses on Differentially private data release for data mining. Partition-based approach divides a given data set into disjoint groups and releases some general information about the groups. The two most popular anonymization techniques are generalization and bucketization. Generalization makes information less precise while preserving the truthfulness of information. Unlike generalization, bucketization does not modify the QID and the sensitive attribute (SA) values but instead de-associates the relationship between the two. However, it thus also disguises the correlation between SA and other attributes and, therefore, hinders data analysis that depends on such correlation. Many algorithms have been proposed to preserve privacy, but only a few have considered the goal for classification. All these algorithms adopt k anonymity or its extensions as the underlying privacy principle and, therefore, are vulnerable to the recently discovered privacy attacks . More discussion about the partition-based approach can be found in a survey paper. Differential privacy has received considerable attention recently as a substitute for partition-based privacy models.

Arik Friedman and Assaf Schuster defines The algorithms that ensure differential privacy by adding noise to the outcome of the logistic regression model or by solving logistic regression for a noisy version of the target function. Unlike the approach considered in this paper, the algorithms require direct access to the raw data. The use of synthetic datasets for privacy preserving data analysis can be very appealing for data mining applications, since the data miner gets unfettered access to the synthetic dataset. Initial results suggest that ensuring the usefulness of the synthetic dataset requires that it be crafted to suit the particular type of analysis to be performed.

3. Existing System

Privacy becomes a more and more serious concern in many applications. One of the privacy concerned problems is publishing microdata for public use which has been extensively studied recently. A large category of privacy attacks is to re-identify individuals by joining the published table with some external tables modeling the background knowledge of users. To battle this type of attacks, the mechanism of k-anonymity was proposed.A data set is said to be k-anonymous (k 1) if, on the quasi-identifier attributeseach record is indistinguishable from at least k1 other records within the same data set.The larger the value of k, the better the privacy is protected.

To solve the k-Anonymization problem for a transactional database, generalization is in use. If original database D does not satisfy the k-anonymity then it is transformed to D' by replacing items with their generalized ones.Here in supermarket database while entering the item is provided with its respective generalization.Generalization replaces initial attribute with generalized attribute.

A. Optimal Anonymization

To find the optimal cut i.e. no generalization that satisfies k m –anonymity and has the least information loss, we can examine systematically the generalizations in the cut hierarchy, in a bottom-up, breadth first fashion. Initially the cut Cng which corresponds to no generalization is put to queue Q. While Q is not empty, we remove first cut from C from it and examine whether it satisfies k m anonymity. If it satisfies then it becomes a candidate solution. If it does not satisfy k m -anonymity, its immediate ancestors in the hierarchy, which do not have a descendant cut that satisfies k m-anonymity are added to the queue.

4. Implementation Details

A. System Overview

The following Figure shows the proposed system architecture.

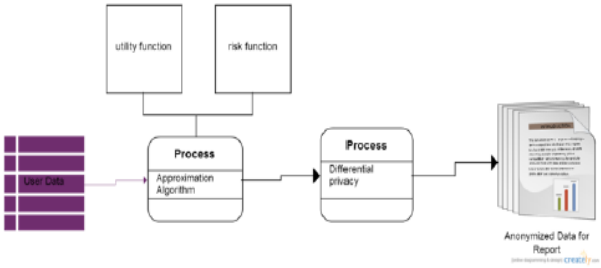


Figure: System Architecture

Licensed Under Creative Commons Attribution CC BY

Above figure explain the architecture of our system.It explain the input, output and function of our system. In above diagram User Data is input to system and Anonymized Data is output.

B. Problem Modeling

1. Let 'S' is the System
S= {I,F,O}
2. Inputs (I) for the System 'S' can be identified as I1, I2, I3, and I4 ..
I= {a,D}
where
a ->Record a ϵ D (Non Anonymized Data from User)
D ->D is Complete Dataset
3. Output (O) is set of outputs System 'S' can provide
O= {d,g,t}
where
d ->Data with proper Anonymization
g -> Graph
t ->Time to complete Algorithms
4. System uses so many Function (F) to generate Output(O) from given Inputs(I) as F1,F2,F3 which can be denoted as
F= {A,L,R,c1,c2,c3}
Where
R ->The risk function
L -> The utility function
A ->An Approximation Algorithm
C1 -> Complete cover
C2 -> Utility maximization
C3 -> differential Privacy

To achieve the high security when the risk threshold is very small is a challenging task.so that the system implements two algorithm .1.Approximation 2.Differential Privacy.Follewing functions are defined by the algorithms:

Utility Function A utility function assigns numerical values ("utilities") to outcomes, in such a way that outcomes with higher utilities are always preferred to outcomes with lower utilities.

A utility function : X-> R represents a preference relation $\preccurlyeq$ on X if for every x,y ϵ X, $u(x) < u(y)$ implies $x \prec y$. If u represents $\preccurlyeq$ then this implies $\preccurlyeq$ is complete and transitive, and hence rational. Risk Function A loss function or cost function is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event. An optimization problem seeks to minimize a loss function. An objective function is either a loss function or its negative (sometimes called a reward function or a utility function), in which case it is to be maximized.

Differential Privacy
Differential privacy aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records.

Consider a trusted party that holds a dataset of sensitive information (e.g. medical records, voter registration information, email usage) with the goal of providing global, statistical information about the data publicly available, while preserving the privacy of the users whose information the data set contains. Such a system is called a statistical database. The notion of indistinguishability, later termed Differential Privacy, formalizes the notion of "privacy" in statistical databases.

5. Experimental Results

A. Experimental Setup

We use an experimental setup.we conducted our experiments on the item description table of Wal-Mart database. The table contains more than 400,000 records each with 30 attributes. The risk components are computed based on both identifiability and sensitivity.

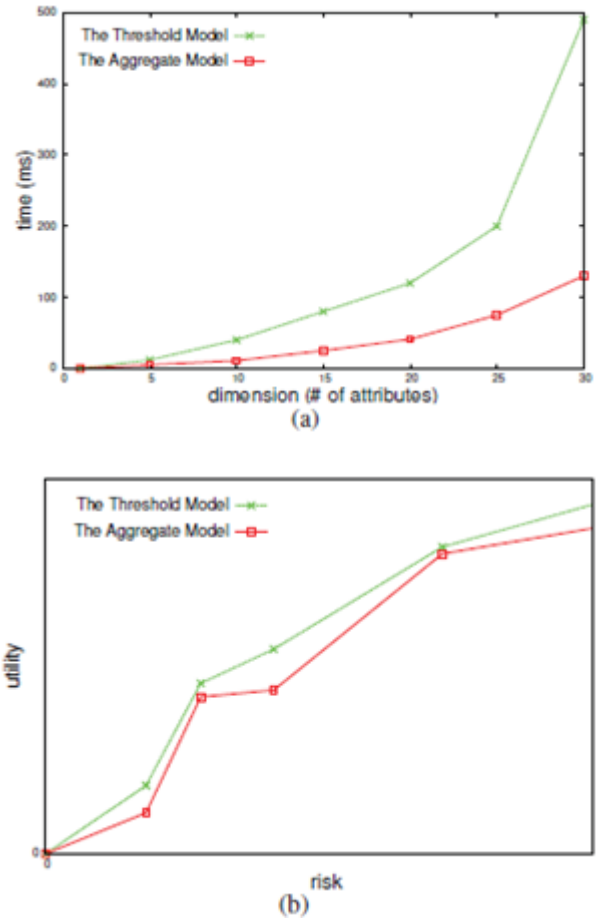


Figure: Impact of Differential Privacy on both (a)Efficiency and (b)risk.

6. Conclusions

Implement an approximation algorithm that computes a nearly optimal solution when the risk threshold is low enough. Also proposed a scalable algorithm that meets differential privacy by applying a specific random sampling. While this might shed some light on the difficulty of obtaining an optimal solution for the threshold model, it may be also possible to extend some of the techniques used for the densest subgraph problem to our problem.

Licensed Under Creative Commons Attribution CC BY

7. Acknowledgment

I would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. I am thankful to the authorities of Savitribai Phule conference, organized by, for their constant guidance's and support. I am also thankful to the reviewer for valuable suggestion. I am also thank the collage authorities for providing the required infrastructure and support. Finally, I would like to extend a heartfelt gratitude to friends and family member.

References

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In Procededings of the International Conference on Very Large Data Bases (VLDB), pages 901909, 2005.
[2] C. Dwork. Differential privacy. In ICALP, pages 112, 2006.
[3] C. Dwork. Differential privacy: A survey of results. In Proceedings of the International Conference on Theory and Applications of Models of Computation (TAMC),pages 119, 2008.
[4] M. R. Fouad, K. Elbassioni, and E. Bertino. Towards a differentially private data anonymization. Technical Report CERIAS 2012-1, Purdue University, 2012.
[5] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In Proceedings of the IEEE International Conference on Data Engineering (ICDE), pages 205216, 2005.
[6] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis. Fast data anonymization with low information loss. In Proceedings of the International Conference on Very Large Data Bases (VLDB), pages 758769, 2007.
[7] V. S. Iyengar. Transforming data to satisfy privacy constraints. In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 279288, 2002.
[8] T. Li and N. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the IEEE International Conference on Data Engineering (ICDE), pages106115, 2007.
[9] L. Lovasz and S. Vempala. Fast algorithms for log-concavefunctions:Sampling,rounding, integration and optimization. In Proceedings of the Annual Symposium on Foundations of Computer Science (FOCS), pages 5768, 2006.
[10] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In Proceedings of the IEEE International Conference on Data Engineering (ICDE), page 24, 2006.
[11] N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu. Differentially private data release for data mining. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD 11, pages 493501. ACM, 2011.
[12] P. Samarati and L. Sweeney. Data to provide anonymity when disclosing information. In Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), page 188, 1998.
[13] L. Sweeney. Privacy-enhanced linking. Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations, 7(2):7275, 2005.
[14] X. Xiao and Y. Tao. Personalized privacy preservation. In Proceedings of the ACM Special Interest Group on Management of Data (SIGMOD), pages 229240, 2006.
[15] J. Cao, P. Karras, P. Kalnis, and K.-L. Tan. SABRE: A sensitive attribute bucketization and redistribution framework for t-closeness. Journal on Very Large Data Bases (VLDB), 20(1):5981, 2011.
[16] M. R. Fouad, G. Lebanon, and E. Bertino. ARUBA: A risk-utility-based algorithm for data disclosure. In Proceedings of the VLDB Workshop on Secure Data Management (SDM), pages 3249, 2008.
[17] Charu C. Aggarwal and Philip S. Yu A Condensation Approach to Privacy Preserving Data Mining at Springer-Verlag Berlin Heidelberg 2004.
[18] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke,and Muthuramakrishnan Venkitasubramaniam l-Diversity: Privacy Beyond k-Anonymity ACM Trans. Knowl.Discov.Data 1, 1,march2007.

Licensed Under Creative Commons Attribution CC BY