World Applied Sciences Journal 29 (8): 1026-1030, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.29.08.1407

# Data Security Using L Diversity

*P. Gayathri*

Bharath University, Chennai, Tamil Nadu, India

**Abstract:** The large amount of data is used in corporate and government institution. The sequence of releases is more identify by adversary. So, the user's privacy is violated. To avoid this to use privacy protection techniques are applied here. Privacy preserving serial data publishing on dynamic databases has relied on unrealistic assumptions of the nature of dynamic databases. In many applications, some sensitive values changes freely while others never change. For example, in medical applications, the disease attributes changes with time when patients recover from one disease and develop another disease.

**Key words:** Sequential Background Knowledge · Background Knowledge sensitive values · Posterior Knowledge Sensitive Values · Anonymity

## INTRODUCTION

The generic name for the collection of tools designed to protect computer data is computer security (including access via network). Collection of tools designed to protect data during transmission between computers in the network is network (internet) security. Security services should provide: confidentiality, authentication, non-repudiation, integrity of transmitted data Security mechanisms should consider possible attacks on the security features. Security attack is any action that compromises the security of information owned by an organization Security mechanism is a mechanism that is designed to detect, prevent, or recover from a security attack [1].

Many security mechanisms are based on cryptographic techniques. Security service is a service that enhances the security of the data processing systems and the information transfers of organization. The services are intended to counter security attacks and they make use of one or more security mechanisms to provide the service. Security services are to replicate functions normally associated with physical documents. Documents usually have: signatures, dates. They may be notarized or witnessed, may be licensed. They may need protection from disclosure, tampering, or destruction. Security services are to take into account specific features of electronic documents. It is usually possible to discriminate between an original paper document and a xerographic copy. However, an electronic document is merely a sequence of bits; there is no difference between the 'original' and any number of copies. An alteration to a paper document may leave some sort of physical evidence of the alteration. For example, an erasure can result in a thin spot or a roughness in the surface. Altering bits in the computer memory or in a signal leaves no physical trace. Any 'proof' process associated with a physical document typically depends on the physical characteristics of that document (e.g., the shape of a handwritten signature or an embossed notary seal). Any such proof of authenticity of an electronic document must be based on internal evidence present in the information itself [2].

**Related Work:** Privacy preserving techniques are of two forms commonly named as Micro data anonymity, differential privacy methods. Simply combining these two types of techniques cannot be a solution, since background knowledge can enable new kinds of privacy threats on sequential data releases. Ways of getting background knowledge's are, Incrementally extracted knowledge, mining data-approximately extracting external data's from same domain and finally exploiting domain knowledge (i.e)extracted from scientific literature.

HD-composition is a protection technique for permanent sensitive values. It contributes two roles holder and decoy. Holder is protected by number of decoys. A decoy is usually a person, device or event

**Corresponding Author:** P. Gayathri, Bharath University, Chennai, Tamil Nadu, India.

meant as a distraction, to conceal what an individual or a group might be looking for. HD. Differential privacy is an approach differs from much (but not all!) of the related literature in the statistics, databases, theory and cryptography communities, in that a formal and a domnia privacy guarantee is defined and the data analysis techniques presented are rigorously proved to satisfy the guarantee. The key privacy guarantee that has emerged is differential privacy. Roughly speaking, this ensures that (almost and quantifiably) no risk is incurred by joining a statistical database. In this survey, we recall the definition of differential privacy and two basic techniques for achieving it. We then show some interesting applications of these techniques, presenting algorithms for three specific tasks and three general results on differentially private learning. The randomized function K is the algorithm applied by the curator when releasing information. So the input is the data set and the output is the released information, or transcript. We do not need to distinguish between the interactive and non-interactive settings.

Fast Data Anonymization [3] publishing micro data without revealing sensitive information, leading to the privacy preserving paradigms of k-anonymity. k-anonymity protects against the identification of an individual's record. `-diversity, in addition, safeguards against the association of an individual with specific sensitive information. However, existing approaches suffer from at least one of the following drawbacks: (i) The information loss metrics are counter-intuitive and fail to capture data inaccuracies inflicted for the sake of privacy. (ii)The anonymization process is inefficient in terms of computation and I/O cost.

k-anonymity- Anonymity is nothing but nameless, it is an question that, How do you publicly release a database without compromising individual privacy. K-anonymity guarantees that the data released is accurate. Methods for achieving k-anonymity is suppression and generalization whereas suppression can replace individual attributes with a* and generalization is replacing individual attributes with a broader category eg (Age 26=>Age 20-30).

A variety of information loss metrics have been proposed. The Classification Metric (CM) is suitable when the purpose of the anonymized data is to train a classifier. Each record is assigned a class label and information loss is computed based on the adherence of a tuple to the majority class of its group. However, it is not clear how CM can be extended to support general purpose applications.

The Discernibility Metric (DM), on the other hand, measures the cardinality of the equivalence class. Although classes with few records are desirable, DM does not capture the distribution of records in the QT space. More accurate is the Generalized Loss Metric and the similar Normalized Certainty Penalty (NCP).

Injector Mining [4] in this approach, one first mines knowledge from the data to be released and then uses the mining results as the background knowledge when anonymizing the data. The rationale of our approach is that if certain facts or background knowledge exist, They should manifest themselves in the data and we should be able to find them using data mining techniques [5-7].

One intriguing aspect of our approach is that one can argue that it improves both privacy and utility at the same time, as it both protects against background knowledge attacks and better preserves the features in the data [8]. We then present the Injector framework for data anonymization. Injector mines negative association rules from the data to be released and uses them in the anonymization process.We also develop an efficient anonymization algorithm to compute the injected tables that incorporates background knowledge. Experimental results show that Injector reduces privacy risks against background knowledge attacks while improving data utility. Aspect of the approach is that one can argue that it improves both privacy and utility at the same time, as it both protects against background knowledge attacks and better preserves the features in the data.

Mondrian Multidimensional K-Anonymity, K-Anonymity has been proposed as a mechanism for protecting privacy in microdata publishing and numerous recoding "models" have been considered for achieving kanonymity. This paper proposes a new multidimensional model, which provides an additional degree of flexibility not seen in previous (single-dimensional) approaches. Often this flexibility leads to higher-quality anonymizations, as measured both by general-purpose metrics and more specific notions of query answerability.

Optimal multidimensional anonymization is NP-hard (like previous optimal k-anonymity problems). However, we introduce a simple greedy approximation algorithm and experimental results show that this greedy algorithm frequently leads to more desirable anonymizations than exhaustive optimal algorithms for two single-dimensional models [9]. The greedy algorithm is substantially *more efficient* than proposed optimal k-anonymization algorithms for single-dimensional models. The time complexity of the greedy algorithm is $O(nlogn)$, while the optimal algorithms are exponential in the worst case.

The greedy multidimensional algorithm often produces *higher-quality* results than optimal single dimensional algorithms [10]. The multidimensional recoding would lend itself to creating anonymizations that are useful for building data mining models since the partitioning pattern more faithfully reflects the multivariate distribution of the original data.

Greedy approximation follows problem solving heuristic of making the locally optimal choice at each stage [11]. It achieves many choices, for achieving changes optimal substructure is provided. Minimal generalization is used to protect the microdatas in statistical and tabular forms during information release. This approach is based on k-anonymity.

**System Model**

**Proposed System:** An defending algorithm JS reduce is used along with an technique called l-diversity. The goal of JS-reduce is to create QI-groups whose tuple respondents have similar $RBK_{sv}$ distributions. In proposed system the JS-reduce algorithm is used to protect the privacy. JS-reduce technique relies on a background knowledge revision process that is not tied to a specific inference method. The privacy preserving techniques to capturing non uniform transition probabilities. JS-reduce also enforce k -anonymity and t-closeness, in order to protect against well-known identity and attribute-disclosure attacks.

The JS-reduce defense guarantees that, for each QI-group Q in an anonymized view, the JS divergence of the set of probability distributions $RBK_{sv}$ of respondents of tuples in Q is below j.

| ID | ZIPCODE | NATIONALITY | DISEASE |
|---|---|---|---|
| 1 | 1302 | RUSSIAN | HEART DISEASE |
| 2 | 1698 | AMERICAN | CANCER |

Fig. 3.1: Microdata

| NAME | AGE | GENDER | ZIP | EX-RES | BKsv |
|---|---|---|---|---|---|
| ALICE | 54 | MALE | 1243 | MAM-POS | 0.087 |
| BETTY | 45 | FEMALE | 1567 | CX-NEG | 0.67 |

Fig. 3.2: Background sensitive values

| EX-RES at t1 | EX-RES at t2 | P(t1/t2) |
|---|---|---|
| MAM-POS | CX-NEG | 0.06 |
| CX-POS | MAM-POS | 0.2 |

Fig. 3.3: Sequential Background knowledge

The actual value of threshold j must be chosen according to many domain-specific factors, including the diversity of sensitive values in released views and background knowledge. A challenging issue in this scenario is the protection of users' privacy, considering that potential adversaries have access to multiple serial releases and can easily acquire background knowledge related to the specific domain. This knowledge includes the fact that certain sequences of values in subsequent releases are more likely to be observed than other sequences. For example, it is pretty straightforward to extract from the medical literature or from a public data set that a sequence of medical exam results within a certain time frame has higher probability to be observed than another sequence. Privacy protection approaches can be divided in microdata anonymity and differential privacy methods.

Microdata anonymity works have focused on techniques dealing either with multiple data releases, or with adversary background knowledge, but limited to a single data release. We are not aware of any work taking into account the combination of these conditions. This case cannot be addressed by simply combining the two types of techniques mentioned above, since background knowledge can enable new kinds of privacy threats on sequential data releases.

Quasi-identifier: A set of non-sensitive attributes {Q1,...,Qw} of a table is called a quasi-identifier if these attributes can be linked with external data to uniquely identify at least one individual in the general population.

One example of a quasi-identifier is a primary key like social security number. Another example is the set {Gender, Age, Zip Code} in the GIC dataset that was used to identify the governor of Massachusetts as described in the introduction. Let us denote the set of all quasi-identifiers by QI. We are now ready to formally define k-anonymity.

The l-Diversity Principle, can be derived in two ways. First, it can be derived in an ideal theoretical setting where it can be shown that the adversary's background knowledge will not lead to a privacy breach. Then we will re-derive the l-diversity principle from a more practical starting point and show that even under less-than-ideal circumstances.

l-diversity can still defend against background knowledge that is unknown to the data publisher. Although the arguments in this subsection can be made precise, we will keep our discussion at an intuitive level for the sake of clarity.

**System Architecture:** The system architecture contributes with the overall process of the system, the data which are considered is an background knowledges of an users. The existing system defines only

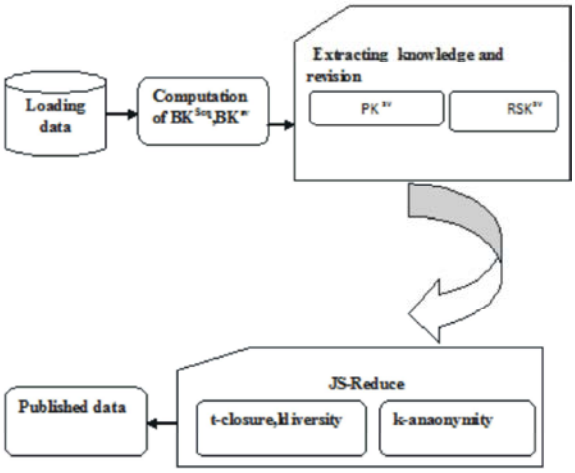


Fig. 3.4: System Architecture

the technique called K-anonymity whereas the below given figure describes the usage of t-closure along with l diversity. k-anonymity provides data quality. l-diversity along with k-anonymity provides security on sequential background knowledge on sensitive data releases.

The microdatas are loaded and the corresponding background sequential and sensitive values are computed accordingly. After that an posterior knowledge($pk^{sv}$) and revised values($RSK^{sv}$) are extracted for secret release of datas.

**Modules Description**

**Extracting Dataset:** The datas are extracted from the database for sequential data release. The datas first extracted are formal information which are said to be raw data. The loading of data is said to be a extraction of data, these datas are furtherly used for computing background sequential knowledge and background sensitive values.

Every medical institution maintained one database. It may contain patient database, worker database and so on. In our project to use patient dataset. The dataset contains name of the patient, age, gender, zip code and medical test result. The dataset are stored in databases. Finally, the data are loaded in project. The loaded data are displayed in table.

**Computation of $BK^{sv}$, $Bk^{seq}$:** In this module is used to computing the values of sensitive background knowledge ($BK^{sv}$). Sequential background knowledge ($Bk^{seq}$). These values are used to compute the $RBK^{sv}$ and $PK^{sv}$ . The sensitive values background knowledge is a function $BK_{sv}$: R -> Ý, where R is the set of possible respondents' identities and is the set of possible probability distributions of S, where D [S] ={$s_1$,$s_2$,$s_3$,…. $S_n$}. Sequential background knowledge is a function that returns the probability distribution of S at $T_j$ given a sequence A=<s1,s2,s3….sj-1> of past observations at T = <t1,t2,….tj-1>.

**Extracting Posterior Knowledge:** The computing posterior knowledge is possible to consider a QI-group at a time. The $PK^{sv}$ is computed according at time t1 and t2. The exact computation of $PK^{sv}$ is intractable. posterior knowledge at $t_i$ represents the adversary's confidence about the association between a respondent and sensitive values after the observation of view Vi*. To denote PKsv at $t_i$ by PK $sv_i$.

**Extracting Revised Knowledge:** Revised sensitive values background knowledge at $t_i$ (i> 1) the adversary needs to calculate, for each respondent r of a tuple in Vi* and for each sensitive value s the marginal probability of r to be the respondent of a tuple with private value s in Vi* given PKsv and BKseq.

The revised knowledge is used to compute the probability of posterior values of sensitive knowledge. The revised knowledge is calculated according to posterior knowledge at time t1 and time t2. The output of revised knowledge is used in JS-reduce algorithm.

**JS-Reduce Defense:** The goal of JS-reduce is to create QI-groups whose tuple respondents have similar $RBK^{sv}$ distributions. If the respondents of tuples in a QI-group are indistinguishable with respect to $RBK^{sv}$. So, adversary cannot exploit background knowledge to perform the attack.

Defending against background knowledge attacks is not sufficient to guarantee privacy protection against other kinds of attacks. For this reason, JS-reduce also enforce k -anonymity and t-closeness in order to protect against well-known identity. Jensen-Shannon divergence (JS) measure is used in to quantify information disclosure.

**Result of Data:** JS-reduce algorithm is used to protect the anonymous data. So, the result of defense is to provide the strong protection and good data quality. To find the groups of tuples satisfying privacy constraints and to limit the generalization of QI values. If the required privacy constraints are satisfied, a new QI-group is created. The QI values are substituted with intervals including the QI values of each tuple and the original tuples are removed from value set.

## CONCLUSION

The privacy protection mechanism is applied here. Our proposed defense algorithm is based on jensen-shannon divergence. This algorithm is to provide the strong protection of data units. The existing system focuses on microdata by enforcing k-anonymity. The sequential background knowledge should be focused for preventing it. The data utility should be considered as main criteria along with data security [12-6].

## REFERENCES


1. Bu, Y., A. Wai, C. Fu, R.C.W. Wong, L. Chen and J. Li, 2008. Privacy Preserving Serial Data Publishing by Role Composition, Proc. VLDB Endowment, 1: 845-856.
2. Dwork, C., M. Naor, T. Pitassi and G.N. Rothblum, 2010. Differential Privacy under Continual Observation, Proc. 42nd ACM Symp. Theory of Computing (STOC '10), pp: 715-724.
3. Ghinita, G., P. Karras, P. Kalnis and N. Mamoulis, 2007. Fast Data Anonymization with Low Information Loss, Proc. 33rd Int'l Conf. Very Large Data Bases, 07: 758-769.
4. Injector Mining Background Knowledge for Data Anonymization-Tiancheng Li, Ninghui Li, J-W Byun, J.Cao, B. Carminati 2010.
5. Mondrian Multidimensional K-Anonymity Kristen LeFevre David J. DeWitt Raghuramakrishnan University of Wisconsin, Madison 2010.
6. Du, Z. Teng and Z. Zhu, 2008. Privacy-MaxEnt: Integrating Background Knowledge in Privacy Quantification, Proc. ACMSIGMOD Int'l Conf. Management of Data (SIGMOD'08), pp: 459-472.
7. Dwork, C., 2006. Differential Privacy, Proc. 33rd Int'l Colloquium on Automata, Languages and Programming (ICALP '06), pp: 1-12.
8. Udayakumar, R., V. Khanna, T. Saravanan and G. Saritha, 2013. Retinal Image Analysis Using Curvelet Transform and Multistructure Elements Morphology by Reconstruction, Middle-East Journal of Scientific Res., ISSN: 1990-9233, 16(12): 1798-1800.
9. Udayakumar, R., V. Khanna, T. Saravanan and G. Saritha, 2013. Cross Layer Optimization For Wireless Network (Wimax), Middle-East Journal of Scientific Res., ISSN: 1990-9233, 16(12): 1786-1789.
10. Thooyamani, K.P., V. Khanaa and R. Udayakumar, 2013. A frame work for modelling task coordination in Multi-agent system, Middle-East Journal of Scientific Research, ISSN: 1990-9233, 15(12): 1851-1856.
11. Thooyamani, K.P., V. Khanaa and R. Udayakumar, 2013. An Integrated Agent System for E-mail Coordination using Jade, Indian Journal of Science and Technology, ISSN: 0974-6846, 6(6): 4758-4761.
12. Nahed, M.A., Hassanein, Roba M. Talaat and Mohamed R. Hamed, 2008. Roles of Interleukin-1 (Il-1) and Nitric Oxide (No) in the Anti-Inflammatory Dynamics of Acetylsalicylic Acid Against Carrageenan Induced Paw Oedema in Mice, Global Journal of Pharmacology, 2(1): 11-19.
13. Panda, B.B., Kalpesh Gaur, M.L. Kori, L.K. Tyagi, R.K. Nema, C.S. Sharma and A.K. Jain, 2009. Anti-Inflammatory and Analgesic Activity of Jatropha gossypifolia in Experimental Animal Models, Global Journal of Pharmacology, 3(1): 01-05.
14. Parmar Namita, Rawat Mukesh and J. Kumar, 2012. Vijay Camellia Sinensis Green Tea. A Review Global Journal of Pharmacology, 6(2): 52-59.
15. Jagadeeswaran, M., N. Gopal, B. Jayakar and T. Sivakumar, 2012. Simultaneous Determination of Lafutadine and Domperidone in Capsule by High Performance Liquid Chromatography, Global Journal of Pharmacology, 6(2): 60-64.
16. Yogeswari, S., S. Ramalakshmi, R. Neelavathy and J. Muthumary, 2012. Identification and Comparative Studies of Different Volatile Fractions from Monochaetia kansensis by GCMS, Global Journal of Pharmacology, 6(2): 65-71.