ISSN:0975-9646

Akshaya.S et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3784-3786

Issues in Data Privacy

Akshaya.S ,Jayasre Manchari.V L ,MohamedThoufeeq.A

Information , Technology
SVCE

Abstract-Managing a data set with sensitive but useful information, such as medical records, requires reconciling two objectives: providing utility to others and respecting the privacy of individuals who contribute to the data set. The earliest privacy preserving technique was encryption. Following it is the k-anonymity where each tuple in a k-anonymized dataset should appear at least k times. A series of recent papers formalizes the notion of differential privacy. A database privatization mechanism (which may be either interactive or non-interactive) satisfies differential privacy if the addition or removal of a single database element does not change the probability of any outcome. This efficient implementation guarantees privacy for all input databases.

General Terms- Sensitivity, Performance, Reliability.

Keywords Anonymized; Differential privacy; Adversary; Synthetic database; Sensitivity.

1. INTRODUCTION

Privacy, which is a value so complex, so entangled in competing and contradictory dimensions, so engorged with various and distinct meanings, that sometimes despair whether it can be usefully addressed at all. **Cryptography** prior to the modern age was effectively synonymous with *encryption*. The originator of an encrypted message shared the decoding technique needed to recover the original information only with intended recipients, thereby precluding unwanted persons to do the same. Modern cryptography is heavily based on mathematical theory and computer science practice; cryptographic algorithms are designed around computational hardness assumptions. A basic principle behind encrypting stored data is that it must not interfere with access control. If access controls are implemented well, then encryption adds little additional security within the database itself. This could lead to securing data which they did not wish to encrypt or failing to encode data which they did wish to protect. As Encryption protects your personal data e.g. bank details, love letters etc. it also protects drug dealers who make deals from having their messages intercepted, terrorists planning attacks. If you forget your passphrase and/or key file then there is almost no chance of recovering your data. Cryptography has many disadvantages: Encryption Does Not Solve Access Control Problem, Encryption Does Not Protect against a Malicious DBA, Encrypting Everything Does Not Make Data Secure. Cryptography's Secure function Evaluation **does not solve the privacy problem completely.** To overcome this we go for **k-anonymity**.

The k-**anonymity** notion requires that when only certain attributes, known as **quasi-identifiers** (QIDs), are considered; each tuple in a k-anonymized dataset should appear at least k times. K-anonymity means each released record has at least (k-1) other records in the release whose values are indistinct over those fields that appear in external data. So, k-anonymity provides privacy protection by guaranteeing that each released record will relate to at least k individuals even if the records are directly linked to external information. A release of data is said to adhere to k-anonymity if each released record has at least (k-1) other records also visible in the release whose values are indistinct over a special set of fields called the quasi-identifier [1]. The quasi-identifier contains those fields that are likely to appear in other known data sets. Therefore, k-anonymity provides privacy protection by guaranteeing that each record relates to at least k individuals even if the released records are directly linked (or matched) to external information. Generalization involves replacing (or recoding) a value with a less specific but semantically consistent value replace individual attributes with a broader category. Suppression involves not releasing a value at all can replace individual attributes with a *.

A relevant problem arises when data stored in a confidential, anonymity-preserving database need to be updated. The operation of updating such a database, e.g., by inserting a tuple containing information about a given individual, introduces two problems concerning both the anonymity and confidentiality of the data stored in the database and the privacy of the individual to whom the data to be inserted are related data are referred is often of interest not only to these individuals, but also to the organization owning the database. Because of current regulations, organizations collecting data about individuals are under the obligation of assuring individual privacy. To solve this problem we had proposed a new concept based on combination of encryption and k-anonymity

Thus the field of **differential privacy** has recently emerged as a leading standard of privacy guarantees for algorithms on statistical databases. Let us see in detail about differential privacy and its branches.

2. DIFFERENTIAL PRIVACY

The cornerstone of the new approach to privacy is the definition of differential privacy, which first appeared in [3]. Intuitively, the definition captures the risk of joining the database, where the risk is measured as the adversary's success in predicting whether a single record is present in the database, given the rest of the database. The definition gives un conditional guarantees (including privacy for (small) groups) against a powerful adversary, preserved by sequential composition, and still al-lows many types of statistical or machine learning analyses, as shown in [4-6].

Dalenius in 1977 defined differential privacy as:

"Anything that can be learned about a respondent from the Statistical database can be learned without access to the database."

Definition 1. *A randomized function $\mathcal{K}$ gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,*

$$\Pr[\mathcal{K}(D_1) \in S] \leq \exp(\epsilon) \times \Pr[\mathcal{K}(D_2) \in S] \quad (1)$$

The probability is taken is over the coin tosses of $\mathcal{K}$.

A mechanism *K* satisfying this definition addresses concerns that any participant might have about the leakage of her personal information: even if the participant removed her data from the data set, no outputs would become significantly more or less likely. A database privatization mechanism (which may be either interactive or non-interactive) satisfies differential privacy if the addition or removal of a single database element does not change the probability of any outcome of the privatization mechanism by more than some small amount. It can be achieved in two ways: **Interactive mechanism and Non-interactive mechanism [7]**

In **non-interactive approach**, [8] We consider private data analysis in the setting in which a trusted and trustworthy curator, having obtained a large data set containing private information, releases to the public a "sanitization" of the data set that simultaneously protects the privacy of the individual contributors of data and offers utility to the data analyst. The sanitization may be in the form of an arbitrary data structure, accompanied by a computational procedure for determining approximate answers to queries on the original data set, or it may be a "synthetic data set"

In the **interactive approach**, only the questions actually asked receive responses. In interactive mechanism, Multiple Queries, are given and the queried sample is matched with the adversary's database to find the individual details

These two approaches can be achieved using the following methods.

Existing methods for satisfying differential privacy includes the following:-

- adding noise according to the global sensitivity
- adding noise according to the smooth local sensitivity

2.1 Sensitivity:

2.1.1 Global sensitivity

Simple framework for output perturbation with strong privacy guarantees (1)Noise levels small enough to allow meaningful analysis (2)General interface

The global sensitivity of f is

$GS_f = \max_{x;y:d(x;y)=1} ||f(x)- f(y)||$

Theorem: If $A(x) = f(x) + Lap(GS_f/ \epsilon)^d$, then A is *!*-differentially private.[9]

Noise distribution

Laplace distribution Lap(**λ**) has density $h(y) \alpha e^{-||y||1/}$ **λ**

2.1.2 Local sensitivity

The local sensitivity of f is

$LS_f = \max_{x': \text{neighbor of } x} ||f(x) - f(x')||$

2.1.3 Smooth bounds on sensitivity

Design sensitivity function S(x)

• S(x) is an ϵ -smooth upper bound on LS f (x) if:

– for all x: $S(x) \geq LS\ f(x)$

– for all neighbours x, x': $S(x) \leq e^{\epsilon} S(x')$

Theorem:

If $A(x) = f(x) + noise(S(x)/ \epsilon)$, then A is ϵ' - indistinguishable

3. LAPLACE MECHANISM

The Laplace mechanism only answers single one-dimensional statistics. We evaluate the privacy and utility performance of Laplace noise addition for numeric data. Our results indicate that Laplace noise addition delivers the promised level of privacy only by adding a large quantity of noise for even relatively large subsets. The Laplacian distribution has been used in speech recognition to model priors on DFT coefficients.

A random variable has a Laplace(μ, b) distribution if its probability density function is

$$f(x|\, \mu, b) = \frac{1}{2b} \exp(-|x- \mu| / b)$$

Here, μ is a location parameter and $b \geq 0$, which is sometimes referred to as the diversity, is a scale parameter. If $\mu = 0$ and $b = 1$, the positive half-line is exactly an exponential distribution scaled by 1/2.

The addition of noise drawn from a Laplacian distribution, with scaling parameter appropriate to a function's sensitivity, to the output of a statistical database query is the most common means to provide differential privacy in statistical databases .Consequently, after just a few queries, the intruder's knowledge gain is so large that differential privacy based Laplace noise addition procedure offers no privacy at all.

4. EXPONENTTIAL MECHANISM

General mechanism that yields differential privacy and it is defined and evaluated by considering all possible answers. Any differential private mechanism is an instance of exponential mechanism [10]. Exponential mechanism is used in case of non-numeric queries. The exponential mechanism E takes in the scoring function score and a dataset A parameter, We start by defining a scoring function score: D * R ->R that takes in a dataset A and output r and returns a real-valued score; this score tells us how "good" this output r is for this dataset A, with the understanding that higher scores are better.

Let D be the domain of input datasets.

Let R be the range of noisy outputs.

E(A; score; ∈) = output r with probability proportional to exp(ϵ/2 score(A; r))

Sensitivity of scoring function*:
$\Delta = \max_{r;A,B \text{ where} |AB|=1} |\text{score}(A;r) \text{ score}(B;r)|$

(1) Receive the query f and the prior knowledge Pf from the database user.(2) Compute the actual value of the query response, f(D).(3) Modify Pf to adjust it to f(D) as much as possible, given the constraints imposed by differential privacy. (4) Randomly sample the distribution resulting from the previous step, and return the sampled value as the response to f evaluated at D.

5. MEDIAN MECHANISM

This mechanism can answer exponentially more queries than the previously best known interactive privacy mechanism (the Laplace mechanism, which independently perturbs each query result). With respect to the number of queries, our guarantee is close to the best possible, even for non-interactive privacy mechanisms. Conceptually, the median mechanism is the first privacy mechanism capable of identifying and exploiting correlations among queries in an interactive setting.[11] The basic implementation of the median mechanism is not efficient, alternative implementation runs in time polynomial in n, k, and j X j, and satisfies the following

for every sequence f1; : : : ; fk of predicate queries, for all but a negligible fraction of input distributions, the efficient median mechanismis (∈;δ)-useful.[12]

6. CONCLUSION

Thus in this paper we have discussed all general data privacy issues from cryptographic encryption techniques, k-anonymity to differential privacy. Differential privacy can be achieved by many techniques including laplace, exponential, matrix, geometric, Gaussian and so. Laplace mechanism is used for numeric queries while exponential is for non-numeric queries. Median mechanism is an instance of exponential mechanism and can be used in both interactive and non-interactive differential privacy. Though there are many methods to tackle the data privacy issues the latest and the best is the median mechanism in interactive differential privacy.

7. FUTURE ENHANCEMENT

We pursue this goal by advancing a recent approach to median mechanism, by first answering a different set of queries (a strategy) and then inferring the answers to the desired workload of queries. Although a few strategies are known to work well on specific workloads, finding the strategy which minimizes error on an arbitrary workload is intractable. We prove a new lower bound on the optimal error of this mechanism, and we propose an efficient algorithm that approaches this bound for a wide range of workloads.

ACKNOWLEDGEMENT

We thank Cynthia Dwork,. Blum A., Ligett K., Roth A, Frank Rosar, Tim Roughgarden and Adam Smith for Many useful discussions.

REFERENCES

1. Ninghui Li, Wahbeh Qardaji, Dong Su On Sampling, Anonymization, and Differential Privacy: Or,*k*-Anonymization Meets Differential Privacy
2. T. Dalenius. Finding a needle in a haystack – or identifying anonymous census record.*Journal of Official Statistics,* 2(3):329-336, 1986.
3. Dwork, C.: Differential privacy. Invited talk. In: ICALP (2). (2006)
4. Dwork, C., Nissim, K.: Privacy-preserving datamining onvertically partitioned databases. In: CRYPTO 2004
5. Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar,K.: Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In: Principles Of Database Systems 2007.273–282
6. Blum, A., Ligett, K., Roth, A.: A learning theory approachto non-interactive database privacy. In: STOC 2008. 609–618
7. Cynthia Dwork and Adam Smith :Differential Privacy for Statistics: What we Know and What we Want to Learn in Journal of Privacy and Confdentiality (2009) 1, Number 2, pp. 135-154
8. A Learning Theory Approach to Non-Interactive Database Privacy by Avrim Blum, Katrina Ligetty, Aaron Roth.
9. Sensitivity-Independent Differential Privacy via prior Knowledge Refinement by jordi soria-comas and josep domingo-ferrer
10. Notes on the Exponential Mechanism. (Differential privacy) Boston University CS 558. By Sharon Goldberg.
11. Continuous decisions by a committee : median versus average mechanisms by Frank Rosar
12. Interactive Privacy via the Median Mechanism by Aaron Roth and Tim Roughgarden