BOOSTING THE PRIVACY OF REAL TIME DATA WITH DIFFERENTIAL PRIVACY

Ms.Neeta Patil[1], Prof.Pankaj Agarkar[2]

[1] *PG Student, Computer Engineering, Dr.D.Y.Patil School of Engineering,Lohgaon,Pune,India*
[2] *Assistant Professor, Computer Engineering, Dr.D.Y.Patil School of Engineering,Lohgaon,Pune,India*

Abstract - *Performing data mining of real time data has great value. Generally it has importance in real time applications like traffic congestion applications, disease surveillance applications. In these types of applications data values are highly correlated so providing differential privacy to this data is essential. Hence while maintaining the confidentiality of data, the purpose of statistical database is to allow third parties or users to study the properties of traffic congestion applications or contents in the databases but individual users contribution is keep safe from damage. Trusted party carry the whole database and provide requested information to third party in the form of count. User or third party make query to the trusted party for accessing some data from trusted server. In the form of aggregate function trusted server or party sends the requested data count to user. This aggregate function is formed as answer to the query plus some random noise. So to provide strong privacy guarantee all appreciable data is examine methodically.*

Key Words: Differential privacy, Geometric mechanism, Exponential mechanism, Laplace perturbation, filtering, sampling

1. INTRODUCTION

While considering some applications sharing of real time data has great value. Because many times such data is private data and needs to maintain privacy of this data. For that purpose real time data is mostly shared in the form of aggregate function which provide count to user. When real time data is shared in the aggregate function form security of that data is maintained properly. Thus to provide strong guarantee about confidentiality of data all appreciable data is examine methodically. Over the past few years, privacy of data is preserved by maintaining some terms. Privacy is provided to databases, the general principles of science(theory) and other kind of data. This privacy is provided by means to increase the correctness of queries from statistical databases while decreasing opportunity of identifying its records (differential privacy). In many data mining applications, real time data is provided in the form of aggregate function to provide the security to private data. Applications where such privacy can be provided are Disease Surveillance, monitoring of road traffic, medical records, voter registration information, and email usage. The goal of this differential privacy is analysis of confidential data by preserving the privacy of data. Consider following examples:

Disease Surveillance: Providers of health care collects data from individual visitors. After that all this collected data is shared with third party, who maintains record of this data properly and securely. This third party has all the collected data. So third party may share this data to other parties. Other parties are may be trusted or not.

Monitoring of traffic: Providers of GPS service collect data from individual users. This data is related to location, speed of vehicle and mobility. Again here providers of GPS service share data with third party which is trusted or not.

In recent years lot of research is done on privacy preserving of whatever data get published. Different terms or concepts are there to preserve the privacy of statistical databases. Among those differential privacy has lots of importance. As different applications like disease surveillance, bank database or other maintain lots of data in their databases. This data needs to collect and then stored. All stored data is maintained properly so that guarantee about privacy is provided. So preserving privacy of secure data is big question in front of different organizations and government. Attacker can do any type of attack to access the data from database. While making access of data available to user, strong privacy guarantee needs to provide. Trusted party done not have attackers background knowledge.

2. LITERATURE SURVEY

2.1 Calibrating Noise

Trusted servers carry the database or private information of individual users. User make query to the trusted server for accessing some data. Trusted server sends the requested data to user in the form of aggregate function.. This aggregate function is formed as answer to the query plus some random noise. So to provide strong privacy guarantee all appreciable data is examine methodically. There are two basic categories for privacy i.e. input and output perturbation: In first approach, data are changed randomly and answer to the user created query is calculated from changed data. In second approach i.e. output perturbation, using real data correct answer to the query is calculated and noisy data is provided.

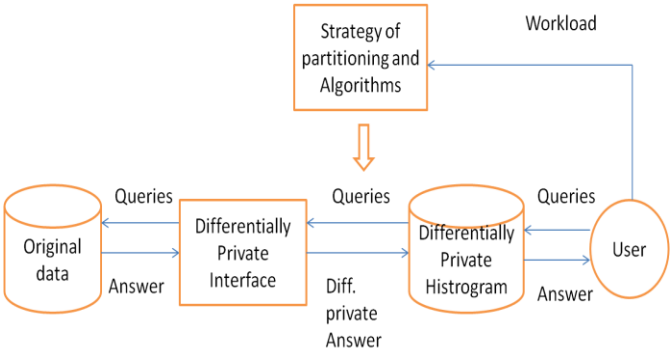


Fig. 2.1 Differentially Private Histogram

2.2 Differentially Private Histogram

Here histogram means it is the partition of database points. Each database points belongs to different partitions. Private access to raw database is given by using the interface Privacy INtegrated Queries platform (PINQ) [6]. Differentially private histogram of raw database is produced by partitioning strategy implementing, which provides queries to the differentially private interface. Histogram then contains the raw database and synthesized dataset. Histogram bear all type of weight of Count queries and other OLAP queries. In this paper[6] two approaches are studied:

1) A baseline strategy
2) A kd-tree partitioning

Kd tree partitioning has original ideas. Foremost in position is the distinctive measure in process of partitioning , in this approach errors are decreased. In second approach, they implement concept with the help of two-step algorithm which produces kd-tree partitions, where histogram generates from cell partitioning.

2.3 ε - Differential Privacy

Xiaokui Xiao[4] specified that several solutions are there to provide privacy guarantee, one of them is differential privacy. To give strong guarantee about privacy concept of differential privacy is used. Those data publishing methods provide differential privacy offer less usefulness. Consider count queries are answered through output dataset, then here noise is equivalent to total number of tuples in the dataset. In this [4],to provide accurate answer to range queries data publishing techniques are studied. Here Privelet data publishing method is presented to give guarantee about differential privacy.

2.4 Private and Continual Statisticsy

In [5], it is specified that as data is continuously updated in the database, the websites provides upgraded data over time while preserving the privacy. Consider different websites, such as online retailers, different search engines like Yahoo, Google and many social networks like facebook, flicker etc. all this provides data in the form of aggregate function. When user make request for any data on search engine like Google, that search engine provide requested data in the form of aggregate function and provides guarantee of data on the server.

2.5 Fast Fourier Transform

Papadimitriou [7] studied to achieve balance between time series compressed property and deviation caused by an outside influence. Based on Fast Fourier Transform and Discrete Wavelet Transform two algorithms are developed so that time series frequencies are disorder. The drawback of this system is addition of noise in data does not give guaranty about differential privacy. It causes attacker or third party which is un-trusted, can get knowledge about stored data on trusted server.

3. SYSTEM ARCHITECTURE

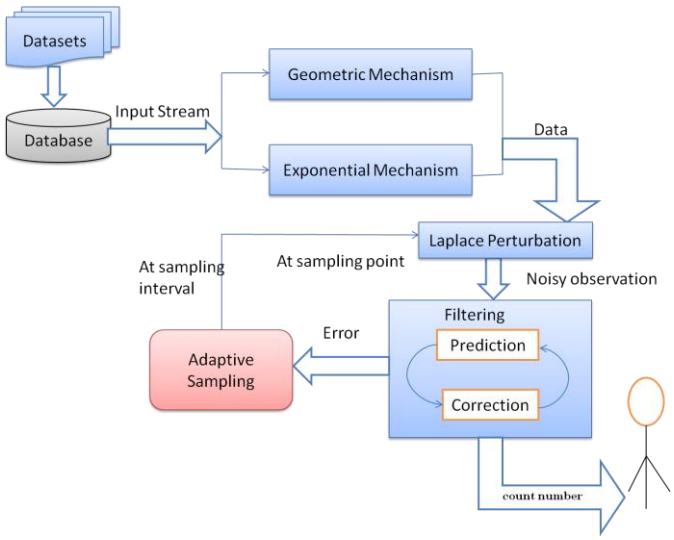


Fig. 3.1 System Architecture

By adding some random noise to the output, exponential mechanism achieves differential privacy. Exponential mechanism is the most common method to do this. Most of differential privacy revolved around real valued functions which have relatively low sensitivity to change in the data of a single individual and by small additive perturbations its usefulness is not hampered. A natural question is what happens in the situation when one wants to preserve more general sets of properties. The Exponential Mechanism helps to extend the notion of differential privacy to address these issues.

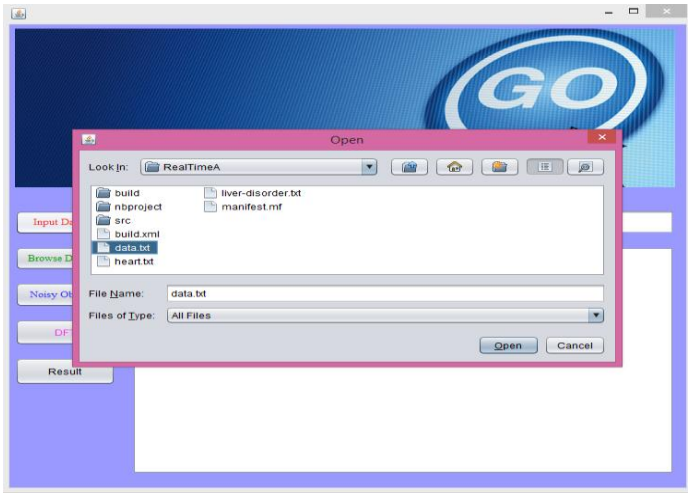
Figure 3.1 shows a block diagram summarizing the main concept of FAST system. In [8], FAST framework is explained. The key steps are as below:

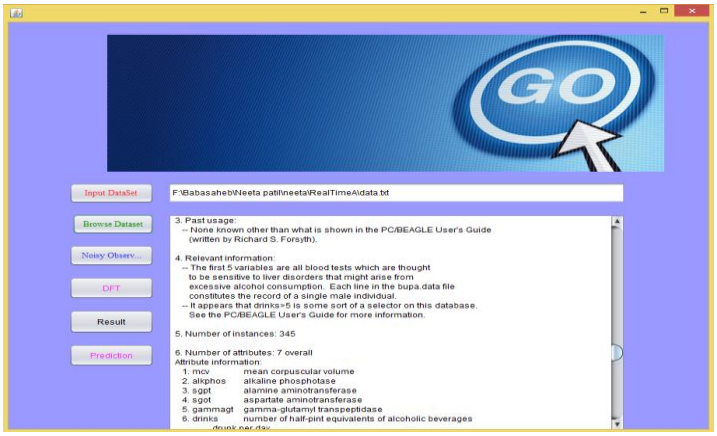
- For each time stamp Ts, the adaptive sampling components determines, whether to query the input time series or not.
- If Ts is sampling point, the data value at Ts is perturbed by the Laplace mechanism to guarantee differential privacy. This value is then pass to filtering module for further process.

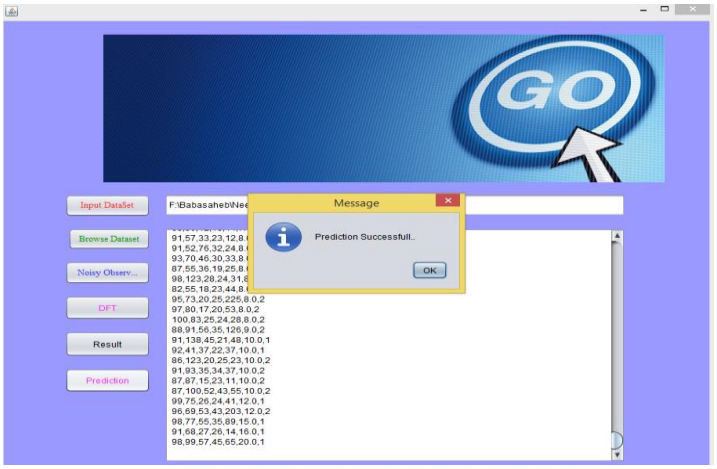
- The filtering module outputs the predicted value at every time stamp. The prediction i.e prior estimate is released to output at a non-sampling point, while a correction i.e. posterior estimate based on both the noisy observation and the prediction, is released at a sampling point.
- The error between the prediction and correction is the pass to sampling component to adjust the sampling rate.

4. RESULTS

After implementing privacy providing system in Java with JSC(Java Statistical Classes) following results are seen. After applying geometric mechanism and exponential mechanism privacy and utility is expected to be increased.







5. CONCLUSIONS

FAST framework provides privacy guaranty using Laplace perturbation. But as privacy guaranty is provided using Laplace Perturbation it minimizes the utility of data. Using Geometric mechanism both utility and privacy is achieved.

ACKNOWLEDGEMENT

I wish to express my sincere thanks to the guide and PG Coordinator Prof. Pankaj Agarkar and Head of Department, Prof. Soumitra Das , as well as our principal Dr.S.S.Sonavne and last but not least, the departmental staff members for their support.

REFERENCES

[1] A. Blum, K. Ligett, and A. Roth, “A learning theory approach to non-interactive database privacy,” in Proc. 40th Annu. ACM STOC, New York, NY, USA, 2008.

[2] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in Proc. 3rd TCC, New York, NY, USA, 2006, pp. 265–284.

[3] M. Hay, V. Rastogi, G. Miklau, and D. Suciu, “Boosting the accuracy of differentially private histograms

through consistency,"PVLDB, vol. 3, no. 1–2, pp. 1021–1032, Sept. 2010.

[4] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," IEEE Trans. Knowl. Data Eng., vol. 23, no. 8, pp. 1200–1214, Aug. 2011.

[5] T.-H. H. Chan, E. Shi, and D. Song, "Private and continual release of statistics," in Proc. 37th ICALP (2), Bordeaux, France, 2010, pp. 405–417, LNCS 6199.

[6] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong, "Publishing set-valued data via differential privacy," PVLDB, vol. 4, no. 11, pp. 1087–1098, Aug. 2011.

[7] S. Papadimitriou, F. Li, G. Kollios, and P. S. Yu, "Time series compressibility and privacy," in Proc. 33rd Int. conf. VLDB, 2007, pp. 459–470.

[8] Liyue Fan, and Li Xiong, \lq \lq Adaptive approach to real time aggregate monitoring with differential privacy,\rq \rq,IEEE transaction on knowledge and data engineering,vol.26,No.9,september 2014, pp. 2094-2106.

BIOGRAPHIES

Ms. Neeta Patil , PG Student at Dr. D.Y.Patil School of Engineering, Savitribai Phule Pune University.

Email- patilneeta.333@gmail.com

Prof. Pankaj Agarkar, Assistant Professor at Dr.D.Y.Patil School of Engineering, Savitribai Phule Pune University.

Email - pmagarkar@gmail.com

