Differentially Private over Infinite Streams

Mian Cheng
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: cm@nudt.edu.cn

Yipin sun
Twentieth Century Fox
Springfield, USA
Email: homer@thesimpsons.com

James Kirk
and Montgomery Scott
Starfleet Academy
San Francisco, California 96678-2391

***Abstract*—With the rapid advances in hardware technology, data streams are being generated daily in large volumes, enabling a wide range of real-time analytical tasks.**
With the data explosion we have been experiencing recently, the problem of ..
However, individual privacy has been a major public concern in data sharing.
Publishing real-time statistics of private data can be of great value when data mining can be performed to understand important phenomena such as traffic congestion or influenza outbreak.
In this paper, we consider the we propose a flexible privacy model of XXX-privacy to enhance

Experiments using real-life datasets show that our proposed algorithms improves the accuracy of and has good performance even under very small privacy cost.
Extensive experiments confirm the efficiency and the scalability of our techniques and demonstrate the quality of the produced anonymizations.
Numerical experiments conducted on real ship trajectories demonstrate that our proposed mechanism can deliver ship trajectories that are of good practical utility.

In this paper, we consider... , we propose a novel for infinite data stream

1. Introduction

The last decade has been the development of the information collected For example, real-time traffic analysis,

——Here are the current problems——
In real life, the size of the data stream is unpredictable, and its length may be infinite, the arrival of a steady stream of data over time.
However, sharing click streams can lead to serious privacy breaches, such as the notorious AOL privacy scandal, even if the identification of individuals(e.g. IP address or name)
The current state-of-the-art paradigm for privacy-preserving data publishing is differential privacy, which requires that the aggregate statistics reported by a data publisher be perturbed by a randomized algorithm A, so that the output of A remains roughly the same even if any single tuple in the input data is arbitrarily modified. Given the output of A, an adversary will not be able to infer much about any single tuple in the input, and thus privacy is protected.
Releasing anonymized and aggregate information is not safe because of unforeseen background knowledge.It is hard to model the attacker's background knowledge in this big data.
We should propose a privacy model based on differential privacy which make no assumption on the background knowledge of the attacker.
A straightforward application of differential privacy mechanism which adds a Laplace noise to each aggregate value at each time stamp can lead to a very high perturbation error due to the composition theorem.
The goal of our work is to enable the publisher to share useful aggregate statistics over individual users continuously (aggregate time series) while guaranteeing their privacy.

Question: How can data aggregators continually release updated statistics, and meanwhile preserve each individual users privacy and Ł?

However, existing privacy preserving models mainly depend on the background knowledge of adversaries who may harm the data by re-identifying it.
Differential privacy can protect the presence of an individual regardless of the adversarys background knowledge and give the demarcation of adversarys ability. Differential privacy does guarantee privacy against intrusion by any adversary when all the entities in the database are independent.
The curator may publish the result on a web site once, and multiple queriers may extract any desired information from the released data (eg., a subset of counts) Because the original data may contain sensitive information about individuals, publishing it without any changes may compromise privacy.
Time series:The production of big personal data in the form of sequences of time-stamped real values, called time-series, stores in the personal devices of the individuals.
———what is the relate work ——- Previous work on differential privacy of streaming data mainly focuses on event-level privacy on finite or infinite streams [13], [14], [11], and user-level privacy on finite streams.

——explain user-level and event-level——
FAST takes as input the total amount of timestamps T, which leads to inapplicability in our infinite scenario.
w-event privacy was proposed to make a nice balance between two former privacy definitions.
We strongly believe that there are no universal sanitization solutions that fit all applications, i.e., provide good accuracy in all scenarios.[2]
———what is our novel approach ——-
In this paper, we consider... Towards this end, we introduce a novel schemes
How to measure whether spending privacy budget is a worthwhile investment or not?
It is important to develop an approximation strategy cooperating with dynamic budget allocation component to raise data utility.
Instead of directly adding noise to a real statistics, we function by transformation of original data or a query structure to achieve better overall utility.
The goal of our work is to .
———what is our contribution? ————
Contribution: The key challenge is to improve the utility of the mechanism while preserving privacy level.
In summary, we make the following contributions:
We design an Algorithm to dynamically add noise at each timestamp.
——Here are the session arrangement——
The remainder of this paper is organized as follows. In Section 2, we discuss studies to our own. In Section 3, we cover related work, follow by session 4, where we describe the hardware architecture design. We then...

2. Preliminaries

Write something here...

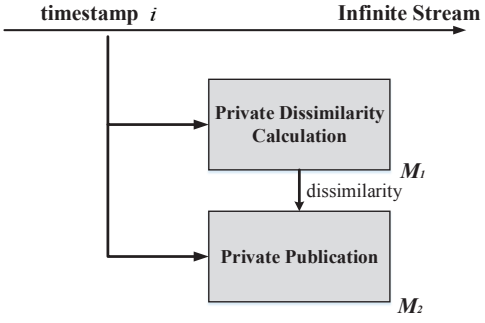

Figure 1. Internal mechanics of M_i

2.1. Differential Privacy

In 2006, Dwork et.al defined the notion of differential privacy. The goal is to publish statistics computed on dataset D, without compromising the privacy of the respondents. Irrespective of whether or not an individual is present in the data set.
Differential privacy has recently emerged as the de facto standard for private data release. This makes it possible to provide strong theoretical guarantees on the privacy and utility of released data.
How to publicly release statistical information about a set of people without compromising he privacy of any individual.
Formally ε-differential privacy is defined as follows:
Definition 1(Differential Privacy) privacy mechanism A gives ε-differential privacy if for any database D and D^* differing on at most one record, and for any possible output $O \in Range(A)$

$$\frac{Pr[A(Q(D))]}{Pr[A(Q(D^*))]} \leq e^{\varepsilon}, \varepsilon > 0$$

Then A achieves ε-DP, $\varepsilon$is a given positive parameter, called privacy budget, used to control unitary privacy level.
Definition 2(Laplace Mechanism)[2]
The first and most widely used method for achieving differential privacy is the Laplace mechanism[], which adds random noise following the Laplace distribution to the true answers to the query functions.
ε-differential privacy can be achieved by adding independent Laplace random noise x to the answer of query Q. In the following, n and D denote the noisy data and a given database, respectively.

$$n = Q(D) + x$$
$$x \sim (x|\lambda) = \frac{1}{2\lambda} * exp(-|x|/\lambda)$$

where λ is a scale parameter of Laplace distribution, which equals to $GS(Q)/\varepsilon$.
Definition 3(Global Sensitivity)
For $f : D \rightarrow \mathbb{R}^d$, the global sensitivity of f is $GS_f = max_{D \sim D'} \parallel f(D) - f(D^{'} \parallel_1$
Trust assumptions
We assume that the data owner and the service provider are trusted, ant the analyst is untrusted.

2.2. Utility metrics

We adopt the classical statistical metric of Mean Absolute Error (MAE) as shown below:

$$MAE = \frac{\sum_{i=1}^{n} |p_i - q_i|}{N}$$

3. Proposed Methods

4. Experimental Evaluation

We have implemented XXX in Java, and used Java Statistical Classes (JSC) to simulating the Laplace distribution.

We experimented with two real datasets. First, we in....
Second, we use a well-known archived dataset named World Cup. It contains 1,352,804,107 Web server logs from the FIFA 1998 Soccer World Cup website, gather in 88 consecutive days[1].
In order to compare the average case, each algorithm runs 50 times then outputs the average results.

5. Conclusion

We have proposed XXX. The key innovation is that our approach...
Publishing accurate trajectories may cause serious privacy breach since such data reveals movement behavior.

Acknowledgments

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.