ε-Differentially Pan-Private Aggregates on Unbounded Streams

Nina Budaeva, David Ding, Angela Gong,
Theresa Lee, Mike Qian, and Kalpana Suraesh
California Institute of Technology
{*nbudaeva,ding,angela,tylee,mqian,ksuraesh*}*@caltech.edu*

Facilitator: Akshay Pai, *appai@caltech.edu*

May 9, 2013

Contents

1 Introduction

Differential privacy and pan-privacy currently is a new area of computer science research that is relatively unexplored. Currently much of the research on privacy mechanisms for databases focuses on static databases. Very few papers talk about stream databases, let alone possible privacy-preserving mechanisms for such databases. The ones that do currently only discuss a private counting mechanism, so we aim to widen the scope with more varied mechanisms. As big data becomes more popular, it is becoming more important to ensure privacy guarantees when dealing with database.

As of this writing, we have not found published mechanisms that provably preserve ε-differential privacy (ε-DP) on unbounded streams for aggregate operators such as sum and average, which are important when dealing with potentially sensitive user data, for example. Therefore, a potential research subject would be to make ε-differentially private and pan-private mechanisms for sum and average. To do so, we would first start with proofs of privacy for previously established mechanisms, and build off of that. Along the way, we might be able to improve upon these mechanisms, such as finding stricter lower bounds or generalizing the mechanisms.

After the completion of our project, we hope to have a wide selection of privacy-preserving stream operators that could one day be used on a practical privacy-aware stream database. We hope that we will be able to outline our algorithms in such a way that it would be easy to translate into code.

2 Related Work

There have been numerous papers regarding the theory and implementation of pan-privacy and continual release of statistics. Here we aim to highlight and summarize the most significant research relevant to our topic.

2.1 Differential Privacy Under Continual Observation

In [3], Dwork *et al.* discuss the concepts of pan-privacy and streaming, and compares it to the definitions of ε differential privacy. The paper describes a pan-private counter and determines its privacy guarantees. In particular, the paper goes over the event-level counter algorithm and shows that it satisfies pan-privacy as defined before in the paper, and that a general event-level counter has a logarithmic lower bound in error. There is also further discussion of transforming a generic single-output streaming algorithm into one that continually produces output, which may be useful for our research.

2.2 Private Decayed Predicate Sums on Streams

Bolot *et al.* discuss bounded mechanisms for various sums on streams in [1]. Specifically, they include mechanisms and algorithms for normal sum (called windowed sum here), which can be reduced to computing the difference of two running sums; polynomial decay sum, which wants to accurately estimate the sum of data which decays; and the exponentially decaying sum, which is similar to the polynomial decay sum. They also address the decayed histogram problem as well as lower and upper bounds on errors for these mechanisms.

2.3 Private and Continual Release of Statistics

In [2], Chan *et al.* discuss the concepts of pan-privacy and consistency, introducing concepts such as p-sums, which are intermediate results from which an observer can estimate the count at every time step. They also provide various time-bounded mechanisms for outputting differentially private counts, which generally become more complex as the error bounds get tighter. The most useful of these is the Binary Mechanism, which uses a noisy binary frequency tree to compute counts. A major contribution of this paper is a method to convert any bounded ε-differentially private mechanism into an ε-differentially private unbounded mechanism. Unbounded mechanisms do not require prior knowledge of an upper bound on the time for which the mechanism will run. Even when it runs for an indefinite amount of time, these unbounded mechanisms provide pan-privacy, differential privacy, and usefulness guarantees.

3 Features

3.1 Definitions and Notation

We aim to have our notation consistent with that used in [3]. The following notation will be used throughout this proposal and our project.

3.1.1 Differential Privacy in the Traditional Setting

A mechanism's privacy is formally defined in terms of differential privacy, as first discussed in [3]. A mechanism is considered differentially private if the output is indistinguishable when run on two nearly identical input streams. An adversary who is trying to take the data will be unable to figure out whether or not an event took place by looking at the results of the mechanism.

Two input streams σ and σ' are adjacent if they differ by at most one tuple t. A mechanism $\mathcal{M}$ preserves ε-DP if for any adjacent streams σ and σ', and for any subset of the possible outputs of the mechanism $S \subseteq Range(\mathcal{M})$,

$$Pr[\mathcal{M}(\sigma) \in S] \leq \exp(\varepsilon) \cdot Pr[\mathcal{M}(\sigma') \in S]$$

3.1.2 Pan-Privacy

What we are most interested about is pan-privacy, which is an extension of differential privacy. A pan-private mechanism is one that can preserve differential privacy even if an adversary is able to access the intermediate states of the stream. Pan-privacy guarantees are such that even if the government or some other organization access the database in between or during calculations, the information they gather is also noisy and does not reveal any private information.

3.1.3 ε-Differential Pan-Privacy

Since we want to do both what was described in Section 3.1.1 and Section 3.1.2, we would like to define an ε-differentially pan-private mechanism to be a mechanism that simultaneously preserves both ε-DP and pan-privacy. Such a mechanism is said to preserve ε-differential pan-privacy.

3.2 Deliverables

To summarize, we hope to achieve the following:

- Sum (unbounded)
- Average (bounded and unbounded)
- Count for non-binary data (where data is not just a 1 or 0)
- Count for bursty streams (where more than one data point may occur at each time step)

If time permits:

- Variance (bounded and unbounded)

The target end results of this project are unbounded mechanisms for computing privacy-preserving sums and both bounded and unbounded mechanisms for computing privacy-preserving averages. We also aim to have general algorithms that can do these mechanisms, as well as proofs that they do, in fact, preserve ε-differential pan-privacy. We want to ensure that they still provide some optimal usefulness. We will use the standard definition of ε-DP, as detailed in Section 3.1.

For sum, we plan on modifying the proofs of count as described in the papers of Section 6. Perhaps one way to look at it is to consider counts over intervals rather than time steps. Non-binary count is just an extension of sum after this. Variance should be approachable in a similar manner to sum, once that has been accomplished, since that is also simply a transformation applied to the incoming value and partial result. As most of the papers we have read seem to follow this approach of taking [3] as a baseline and moving forward using similar methods, we believe this method should work.

4 Project Plan

4.1 Milestones

Our project will span approximately four weeks, from May 10 to June 10, 2013. It is split into four milestones, which are detailed below. Each milestone will take about one week to reach.

4.1.1 Milestone 1: Research Current Algorithms

Our first milestone is to research current papers dealing with pan-privacy and streaming algorithms. We will summarize the main algorithms in each paper and select the best ones to improve upon. We will start with [1-3] and continue from there. Every person will do research and read at least two different papers.

4.1.2 Milestone 2: Invent New Algorithms

After researching current algorithms, we will try to improve upon them and come up with our own algorithms. Our goal is to have around four new algorithms that have not been implemented before. This may require further research to ensure that indeed these algorithms are new. Each person in the group will try to come up with their own algorithm, and pair up once we decide on the best ones to try and prove.

4.1.3 Milestone 3: Prove ε-DP and Pan-Privacy

The next step is to prove that validity of the algorithms we come up with in Milestone 2, by showing that they satisfy ε-DP and streamed pan-privacy definitions. Depending on the number of algorithms we decide to prove, there may be one or two people per algorithm. If we are unable to prove privacy, then we may need to explain why such an algorithm or mechanism may never satisfy ε-differential pan-privacy.

4.1.4 Milestone 4: Final Write-up

Upon completion of the proofs, we will need to document our results in our final write-up. This will be done in a form similar to that of [2]. Everyone will contribute to this final document.

4.2 Backup Plan

Should unforeseen circumstances prevent the group from finishing the tasks mentioned in Section 4.1, we have a backup plan. The worst-case scenario is one in which we are not able to come up with a mechanism for sum or average that is differentially private. More likely, even if we come up with a mechanism, we may not be able to provide a strong enough proof showing that it is sufficiently private and/or useful.

If the proofs do not work out, we will write up the tactics we took which were not sufficient to prove ε-differential pan-privacy, so others do not waste time showing the same result in the future.

We may also attempt to analyze exactly what caused the proofs to fail. Understanding such causes may shed light on the more general problem of creating such algorithms. If we find where the proofs of privacy break down, or what causes them to break down, then we may have a good lead on what went wrong in our proposed algorithm. Knowing this may provide a good platform for anyone else attempting to develop the algorithm. In a more extreme case, we may find from analyzing the breakdown of the proofs that privacy preservation is impossible for such aggregates.

5 Evaluation

There will be a number of ways that the instructor can assign credit to each group member. Every member will contribute to all tasks as noted in Section 4.1, which can provide some basis for the amount of work done. The following will also be helpful for the instructor:

- Self evaluations done by each member of the team.
- Group evaluation in which each member evaluates every other member and their contributions.
- Final write-up will contain a discussion on what each team member did and whether or not each member achieved their individual goals for the assignment.
- Quality of content in blog posts.

6 References

[1] J. Bolot, N. Fawaz, S. Muthukrishnan, A. Nikolov, N. Taft. Private decayed predicate sums on streams. In *Proceedings of the 16th International Conference on Database Theory (ICDT '13)*, pages 284-295, 2013.

[2] T-H. H. Chan, E. Shi, D. Song. Private and continual release of statistics. *ACM Transactions on Information and System Security*, 14(3):1-24, 2011.

[3] C. Dwork, T. Pitassi, M. Naor, G. Rothblum. Differential privacy under continual observation. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC 10)*, pages 715-724, 2010.

[4] C. Dwork, M. Naor, T. Pitassi, G. Rothblum, S. Yekhanin. Pan-private streaming algorithms. In *Proceedings of the 1st Symposium on Innovations in Computer Science (ICS 2010)*, 2010.

[5] D. Mir, S. Muthukrishnan, A. Nikolov, R. Wright. Pan-private algorithms via statistics on sketches. In *Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 11)*. ACM, 2011.