Poster: Towards Scaling Privacy Strength

Joshua Joy
University of California - Los Angeles
jjoy@cs.ucla.edu

Mario Gerla
University of California - Los Angeles
gerla@cs.ucla.edu

***Abstract*—In this paper, we introduce a privacy mechanism that improves the privacy strength while preserving utility. That is, we perform query expansion to reduce the information leakage due to an individual's participation.**

I. Introduction

In this paper we examine the question, how to improve the privacy strength while preserving utility. For example, increasing the amount of privacy noise will certainly improve privacy strength. However, there is only so much privacy noise that might be added before the utility is no longer useful.

Say the Laplace mechanism is used. The privacy strength of a given mechanism is determined by epsilon ε value, which corresponds to the privacy loss measured as the ratio of the max difference between any two differing outputs. A large value of ε means that the privacy loss is large, thus requiring a large amount of privacy noise. Naturally, it follows that increasing the value of ε adds privacy noise mitigating any utility benefits.

Another observation is that the use of the Laplace mechanism requires each individual to truthfully respond, relying on the output perturbation to provide privacy. This requires extra caution in the sensitive queries posed. For example, if a query is posed "Did you take your cholesterol medicine today?" yet the query is only presented to those with heart disease, we can easily infer that someone that participates in this study has heart disease. Clearly no matter how much noise is added, straightforward auxiliary information exists.

One approach is to modify the query in question. That is, we perform **query expansion** whereby we transform the majority population into a minority population. For example, following the previous example we query both diabetes population and the healthy population asking "Did you take your diabetes medicine today?" and "Did you take your multi-vitamin today?". Thus, participation no longer directly implies a participating individual has heart disease.

Techniques such as k-anonymity and l-diversity have been proposed as techniques of hiding an individual in a crowd [7], [6], [5]. However, these techniques are vulnerable to privacy inferences against multiple queries or auxiliary information and do not necessarily perform perturbation or add privacy noise.

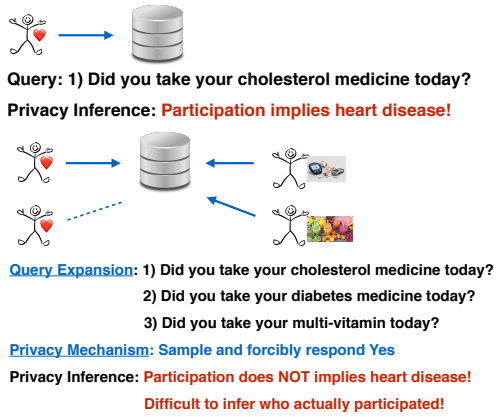


Figure 1. **Query Expansion.** Participation in the first query leaks the individual's medical condition. Participation diversification by query expansion no longer implies a given medical condition.

Thus, the question is how do increase the privacy strength? Increasing the differential privacy noise past a given threshold will mitigate the utility and increasing the population size alone is not sufficient.

However, if we were to perform sampling we could also decrease the inference strength of determining whether or not a given individual participated [3]. Thus, we combine sampling with *increasing* the population size can provide us with strong definitions of privacy.

The problem remains of addressing the utility of such sampling and population size mechanisms. Increasing the sampling rate increases the standard error as adding more individuals who do NOT have the attribute in question simply adds more noise. Techniques such as the randomized response technique and it's variants privately estimate heavy-hitters due to the large standard error [1].

In this paper we present a mechanism that scales in privacy strength yet preserves utility. We perform *query expansion* while we perform sampling across the entire population. Our goal is queries against large populations, while protecting a minority population relying on perturbation by each data owner. We evaluate our scheme over real traffic information collected by the California Transportation Department.

II. Privacy Mechanism

Our query expansion will continue to add the "No" population which means that we must calibrate the sampling rate to avoid incurring a large standard error due to variance.

Network and Distributed Systems Security (NDSS) Symposium 2018
18-21 February 2018, San Diego, CA, USA
ISBN 1-1891562-49-5
http://dx.doi.org/10.14722/ndss.2018.23xxx
www.ndss-symposium.org

Thus, we calibrate the sampling rate standard deviation to the expected population size. This means that when the query expansion is being performed, some effort must be made to estimate the target population size beforehand. However, in cases a meaningful estimate is not able to be performed, the issuer of the query will need to issue a probe query and then issue the calibrated query.

(Round One) In the first round each data owner tosses a biased coin π_{Yes} and π_{No} corresponding to the truthful "Yes" and "No" subpopulations respectively. Heads means that the data owner is forced to respond "Yes", otherwise response "No" for tails. We are careful to calibrate the sampling rate to reduce the standard deviation according to the expected population size due to the query expansion and expected subpopulations (Yes and No).

$$Round\ One_{Yes} = \begin{cases} \mathbf{1} & \textbf{with probability } \pi_{s_{Yes_1}} \\ 0 & \text{with probability } 1 - \pi_{s_{Yes_1}} \end{cases} \quad (1)$$

$$Round\ One_{No} = \begin{cases} \mathbf{1} & \textbf{with probability } \pi_{s_{No_1}} \\ 0 & \text{with probability } 1 - \pi_{s_{No_1}} \end{cases} \quad (2)$$

At this point, privacy noise has been added and thus the underlying truthful distribution is becoming distorted as the number of non-truthful data owners participate. The distortion makes it difficult to estimate the the underlying truthful distribution as we have one equation and two variables (number of truthful and non-truthful data owners).

Thus, we execute a second round. We conduct a fresh sample again being careful to calibrate the sampling rate to minimize the variance for each population enabling us to solve for the truthful population estimate.

(Round Two) In the second round we perform a fresh sample, though allow for the sampling rates to be adjusted if needed.

$$Round\ Two_{Yes} = \begin{cases} \mathbf{1} & \textbf{with probability } \pi_{s_{Yes_2}} \\ 0 & \text{with probability } 1 - \pi_{s_{Yes_2}} \end{cases} \quad (3)$$

$$Round\ Two_{No} = \begin{cases} \mathbf{1} & \textbf{with probability } \pi_{s_{No_2}} \\ 0 & \text{with probability } 1 - \pi_{s_{No_2}} \end{cases} \quad (4)$$

Further details regarding the estimation, privacy guarantee, and private upload can be found in our technical report [4].

III. Evaluation

To examine the utility of our mechanism, we conduct a traffic flow analysis where we privately crowdsource and aggregate vehicles' locations. That is, rather than each vehicle reporting it's exact location, each vehicle privatizes their location.

We utilize the California Transportation Dataset from magnetic pavement sensors collected in LA\Ventura California freeways [2]. There are a total of 3,865 stations and 999,359 vehicles total. We assign virtual identities to each vehicle. Each vehicle privately announces the station it is currently at.

Figure 2 compares the scalable privacy mechanism to the ground truth data over a 24 hour time period with a confidence interval of 95%. We select a single popular highway station. Every vehicle at the station reports "Yes" while every other vehicle in the population truthfully reports "No". For example, we query 1 million vehicles, the 100 vehicles at the given station truthfully respond "Yes", while the remaining 999,900 truthfully respond "No".

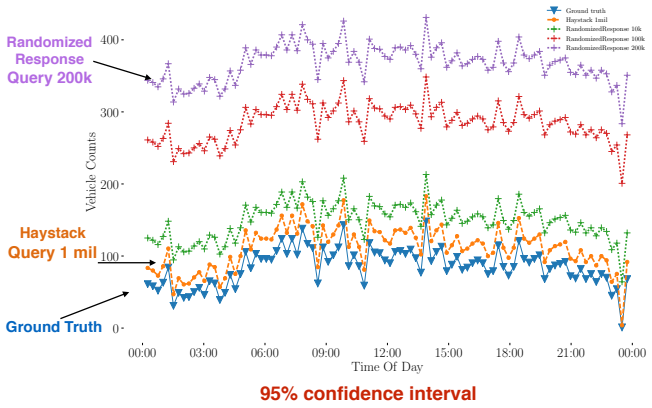


Figure 2. **Estimation Accuracy.** Traffic flow aggregation comparison of ground truth versus privatized vehicle counts with a confidence interval of 95%. Each vehicle privatizes its truthful location and the privacy mechanism aggregates and estimates the underlying traffic flow.

IV. Conclusion

In this paper we have examined the question of increasing the privacy strength while preserving utility. We evaluated our mechanism over actual freeway traffic and demonstrated we can maintain utility even as the participating population increases.

References

[1] A. Bittau, Ú. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnés, and B. Seefeld, "Prochlo: Strong privacy for analytics in the crowd," in *Proceedings of the 26th Symposium on Operating Systems Principles, Shanghai, China, October 28-31, 2017*. ACM, 2017, pp. 441–459. [Online]. Available: http://doi.acm.org/10.1145/3132747.3132769

[2] "California Department of Transportation," http://pems.dot.ca.gov/, 2017. [Online]. Available: http://pems.dot.ca.gov/

[3] J. Gehrke, M. Hay, E. Lui, and R. Pass, "Crowd-blending privacy," in *Advances in Cryptology - CRYPTO 2012 - 32nd Annual Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2012. Proceedings*, ser. Lecture Notes in Computer Science, R. Safavi-Naini and R. Canetti, Eds., vol. 7417. Springer, 2012, pp. 479–496. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-32009-5_28

[4] J. Joy, D. Gray, C. McGoldrick, and M. Gerla, "XYZ privacy," *CoRR*, vol. abs/1710.03322, 2017. [Online]. Available: http://arxiv.org/abs/1710.03322

[5] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *ICDE*, 2007.

[6] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA*, L. Liu, A. Reuter, K. Whang, and J. Zhang, Eds. IEEE Computer Society, 2006, p. 24. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2006.1

[7] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

Towards Scaling Privacy Strength

Josh Joy (jjoy@cs.ucla.edu) and **Mario Gerla** (gerla@cs.ucla.edu)
UCLA Computer Science

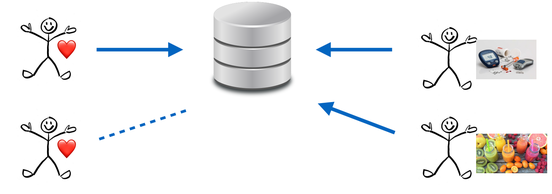
Problem:

- **Increasing the differential privacy noise mitigates utility.**
- **How to improve the privacy strength while preserving utility?**
- **Note: Sampling while increasing truthful "No" population size will increase the standard error.**

Motivation:

Query: 1) Did you take your cholesterol medicine today?

Privacy Inference: Participation implies heart disease!



Query Expansion: 1) Did you take your cholesterol medicine today?
2) Did you take your diabetes medicine today?
3) Did you take your multi-vitamin today?

Privacy Mechanism: Sample and forcibly respond Yes

Privacy Inference: Participation does NOT imply heart disease!

Difficult to infer who actually participated!

Mechanism:

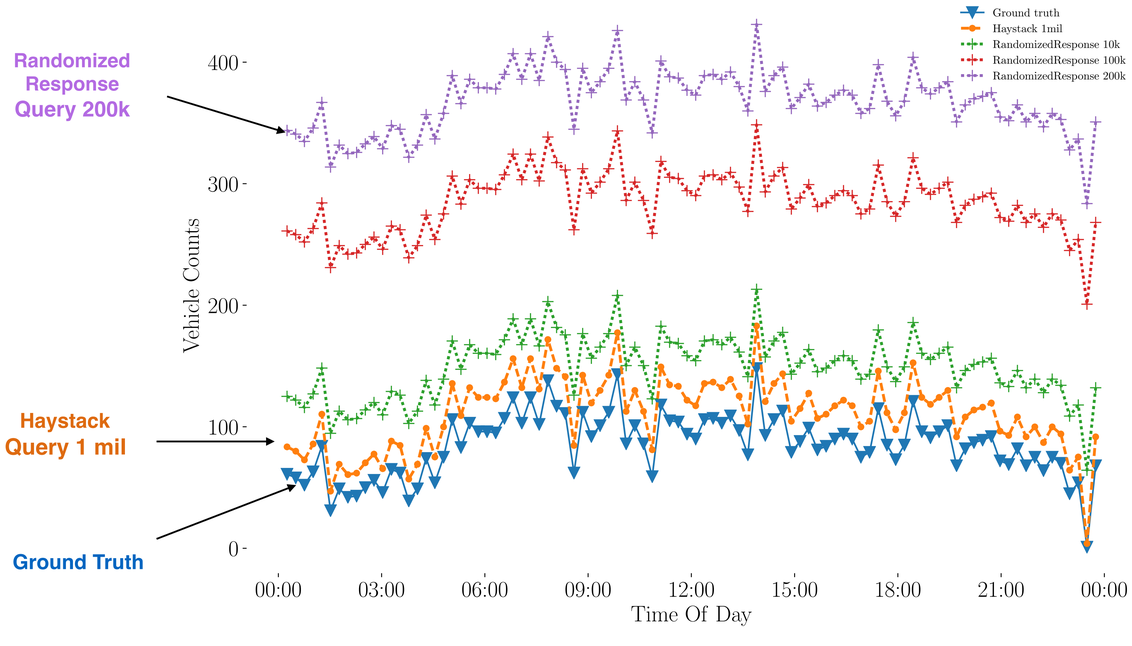
- **Round 1**
  - **Forced Response: Sample truthful "Yes"**
  - **Forced Response: Sample truthful "No"**
- **Round 2 (freshly sample)**
  - **Forced Response: Sample truthful "Yes"**
  - **Forced Response: Sample truthful "No"**
- **Estimation**
  - **1) Aggregate 2) Subtract Round 2 from Round 1 3) Divide by sampling parameter**

Binary Query

Total Population

	Non-Private	Private
Percentage Population	>50%	<50%
Differential Privacy Loss	large ε	small ε

Evaluation:



95% confidence interval