POSTER: Scalable Privacy

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

Dylan Gray
University of California - Los Angeles
dylangray9@cs.ucla.edu

Ciaran McGoldrick
University of California - Los Angeles
ciaran@cs.ucla.edu

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

ABSTRACT

Given that the exact answer to a question is fixed, we ask is it possible to strengthen the privacy by increasing the crowd size that participates even though they do not contribute to the exact answer? In this paper, we introduce the notion of scalable privacy whereby data owners not at a particular location privatize their response such that they respond as if they are at a location (even when they are not). Immediately the question of utility is raised and we examine the tradeoffs to construct such a privacy mechanism so that it scales in both privacy and utility.

CCS CONCEPTS

•Security and privacy → Mobile and wireless security; Privacy protections; •Networks → Network privacy and anonymity;

ACM Reference format:
Joshua Joy, Dylan Gray, Ciaran McGoldrick, and Mario Gerla. 2017. POSTER: Scalable Privacy. In *Proceedings of WiSec '17, Boston, MA, USA, July 18-20, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3098243.3106019

1 INTRODUCTION

In our proposed scalable privacy mechanism, each data owner's sensitive data resides on the data owner's own device. Once receiving a query, each data owner does not directly respond to the query with the sensitive attribute. First, the data owner flips a biased sampling coin to determine whether or not they should participate. If a data owner does not participate, they privately write $\perp$ thus effectively increasing the anonymity set. If a data owner does participate, the data owner locally privatizes their sensitive attribute based on the randomized response mechanism [8] such that only privatized data is released (rather than the sensitive answer). In contrast, prior privacy-preserving systems must synchronize the amount of differential privacy noise added by other data owners or system components [4].

To privately write the data owners' privatized responses to a data aggregator, each data owner generates function secret shares (FSS) [1]. FSS slices the privatized response into multiple shares and transmits one share to each aggregator. Each aggregator independently processes each share within each given epoch. At the end of an agreed upon epoch, all aggregators share their results with the appropriate analyst. As long as there is at least one honest aggregator, the data owners' private write property is guaranteed. FSS hardness assumptions does not depend on a particular pseudorandom number generator (as opposed to a homomorphic pseudorandom generator [5]) which allows the scalable privacy mechanism to be efficient and scalable. Our initial results show we can privately write 250,000 data owners' location data (1280 bits) with a key size of 181KB and an anonymity set of one million data owners.

For example, suppose we crowdsource crowd densities at popular London tourist destinations using the query in Figure 1. A data owner begins by answering the query "Am I at London Bridge?". Prior work using the Laplace mechanism [6, 7] would have everyone at London Bridge answer truthfully. Then, a small amount of privacy noise is added to protect privacy. In Haystack Privacy, all data owners respond to the query as seen in Figure 1. A small fraction of those *not* at London Bridge will respond "Yes, I'm at London Bridge". A small fraction *at* London Bridge will respond "No, I'm not at London Bridge". Both cases provide plausible deniability and are controlled by two different Bernoulli trials specified in the query. To estimate the aggregate count, the expected value of the privacy noise due to the Bernoulli trials is calculated and removed. One observation is the number of people at London Bridge is fixed. While the number of people in any locale (e.g., London Bridge) may be fixed, the inclusion of inputs from people not at that location enables us to increase the crowd size and strengthen the privacy.

2 PRIVACY MODEL

Our system model is in the same setting as local privacy models resembling the randomized response mechanism [8]. However, our model differs in that the output space consists of three values, namely $\perp$, "Yes", and 'No" rather than the two values of "Yes" and "No".

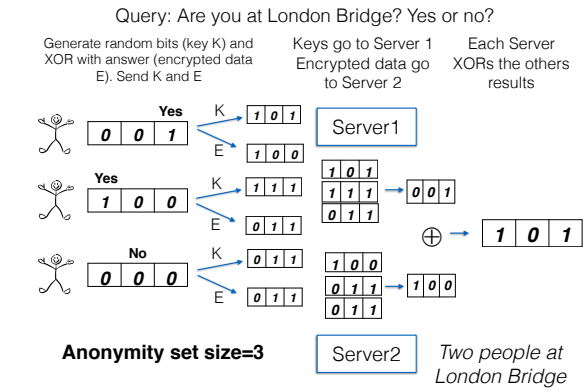
We now define our construction. First, the data owner performs i.i.d. random sampling with probability π_s to determine if they will participate. If not, the data owner writes $\perp$.

Second, the data owner responds with an L bit vector, one bit for each location. The data owner tosses the first coin with probability π_1 for each location. If heads, answer truthfully. If tails, the data owner tosses the second coin with probability π_2 and replies "Yes" if heads and "No" if tails. This means for their actual location, the data owner responds "Yes" with probability $\pi_s \times \pi_1$, and for the locations they are not at, they respond "Yes" with probability $\pi_s \times (1 - \pi_1) \times \pi_2$.

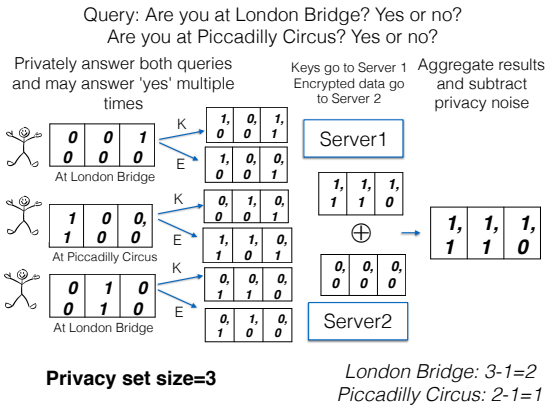
Next, the privatized L bit vector is privately written to the aggregators utilizing the FSS cryptographic primitive. The aggregators verify the shares as described in the previous section utilizing FSS

WiSec '17, Boston, MA, USA

DOI: http://dx.doi.org/10.1145/3098243.3106019



(a) Informaton theoretic private write



(b) Haystack privacy

Figure 1: (a) illustrates that each data owner is able to privately write a value without an adversary linking a message to a particular data owner (b) the Haystack mechanism privatizes the message before privately writing allowing for privacy-preserving aggregate analytics.

share verification. At the end of the epoch, the aggregators finalize the contributed writes into the $N \times L$ matrix without knowing which rows each data owner wrote to.

Definition 2.1. (Scalable Privacy.) Let D be any database. Let a crowd C be defined as any group of at least k data owners.

A privacy mechanism *San* is (k, ε)-haystack-private if for every database D and every data owner $i \in D$, either $San(D) \approx_\varepsilon San(D \setminus i)$, i ε-blends with more than one crowd C, or both.

$San(D) \approx_\varepsilon San(D \setminus i)$ means that essentially removing a specific data owner does not significantly change the result, allowing that data owner to blend in the crowd, such as when sampling occurs.

3 PRELIMINARY RESULTS

We utilize the California Transportation Dataset from magnetic pavement sensors[2] collected in LA\Ventura California freeways [3]. There are a total of 3,865 stations and 999,359 vehicles total. We

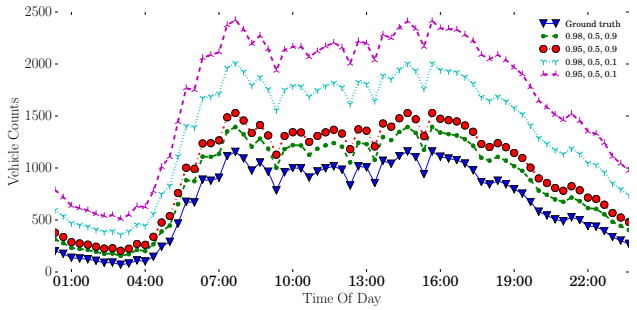


Figure 2: Accuracy. Ground truth versus privatized vehicle counts with a confidence interval of 95%. The legend description refer to the first coin toss π_1, second coin toss π_2, and sampling probability π_s respectively. The Pearson correlation coefficient values are all above 0.9921 with a p-value all lower than 6.8298e−64.

assign virtual identities to each vehicle. Each vehicle 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". The scalable privacy mechanism then privatizes each vehicle's response. The figure shows that the privatized time series is highly correlated with the ground truth. While the L2 error may be large due to the variance, the relative counts is very accurate. Traffic management analyzing the privatized time series would be able to infer the ebbs and flow of the vehicular traffic. The parameters $\pi_1 = 0.998, \pi_2 = 0.5, \pi_s = 0.9$ have the strongest Pearson correlation coefficient with $\rho = 0.9993$ and a p-value of $p = 6.747\text{e}{-100}$. The shape of the privatized line is not the accurate mean estimation though is useful for traffic profile and fluctuations.

REFERENCES

[1] Boyle, E., Gilboa, N., and Ishai, Y. Function secret sharing. In *Advances in Cryptology - EUROCRYPT 2015 - 34th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Sofia, Bulgaria, April 26-30, 2015, Proceedings, Part II* (2015), E. Oswald and M. Fischlin, Eds., vol. 9057 of *Lecture Notes in Computer Science*, Springer, pp. 337–367.
[2] California Department of Transportation. http://pems.dot.ca.gov/.
[3] Google's Waze announces government data exchange program with 10 initial partners. http://www.dot.ca.gov/cwwp/InformationPageForward.do.
[4] Chan, T. H., Li, M., Shi, E., and Xu, W. Differentially private continual monitoring of heavy hitters from distributed streams. In *Privacy Enhancing Technologies - 12th International Symposium, PETS 2012, Vigo, Spain, July 11-13, 2012. Proceedings* (2012), S. Fischer-Hübner and M. K. Wright, Eds., vol. 7384 of *Lecture Notes in Computer Science*, Springer, pp. 140–159.
[5] Corrigan-Gibbs, H., Boneh, D., and Mazières, D. Riposte: An anonymous messaging system handling millions of users. In *2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015* (2015), IEEE Computer Society, pp. 321–338.
[6] Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *TCC* (2006).
[7] Dwork, C., and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science 9*, 3-4 (2014), 211–407.
[8] Fox, J. A., and Tracy, P. E. *Randomized response: a method for sensitive surveys*. Beverly Hills California Sage Publications, 1986.