Poster: Locally Differentially Private Frequent Itemset Mining

Tianhao Wang
Department of Computer Science
Purdue University
West Lafayette, IN
tianhaowang@purdue.edu

Ninghui Li
Department of Computer Science
Purdue University
West Lafayette, IN
ninghui@cs.purdue.edu

Somesh Jha
Department of Computer Science
University of Wisconsin-Madison
Madison, WI
jha@cs.wisc.edu

In recent years, differential privacy [8], [9] has been increasingly accepted as the *de facto* standard for data privacy in the research community. In the standard (or centralized) setting, a data curator collects personal data from each individual, and produces outputs based on the dataset in a way that satisfies differential privacy. In this setting, the data curator sees the raw input from all users and is trusted to handle these private data correctly.

Recently, techniques for avoiding a central trusted authority have been introduced. They use the concept of Differential Privacy in the Local setting, which we call LDP. Such techniques enable collection of statistics of users' data while preserving privacy of participants, without relying on trust in a single data curator. For example, researchers from Google developed RAPPOR [10], [11] and Prochlo [6], which are included as part of Chrome. They enable Google to collect users' answers to questions such as the default homepage of their browser, the default search engine, and so on, in order to understand the unwanted or malicious hijacking of user settings. Apple [16], [17] also uses similar methods to help with predictions of spelling and other tasks. Samsung proposed a similar system [14] which enables collection of not only categorical answers but also numerical answers (e.g., time of usage, battery volume), although it is not clear whether this has been deployed by Samsung. Firefox [1] is also planning to build a "RAPPOR-like" system that collects frequent homepages.

We assume that each user possesses an input value $v \in D$, where D is the value domain. A party wants to learn the distribution of the input values of all users. We call this party the *aggregator* instead of the data curator, because it does not see the raw data. Existing research [4], [10], [18] has developed multiple frequency oracle (FO) protocols, using which an aggregator can estimate the frequency of any chosen value $x \in D$. In [15], Qin et al. considered the setting where each user's value is a set of items $\mathbf{v} \subseteq I$, where I is the item domain. Such a set-valued setting occurs frequently in the situation where LDP is applied. For example, when Apple wants to estimate the frequencies of the emoji's typed everday by the users, each user has a set of emoji's that they typed [17]. The LDPMiner protocol in [15] aims at finding the k most frequent items and their frequencies.

This problem is challenging because the number of items each user has is different. To deal with this, a core technique in [15] is "**padding and sampling**". That is, each user first pads her set of values with dummy items to a fixed size ℓ, then randomly samples one item from the padded set, and finally uses an FO protocol to report the item. When estimating the frequency of an item, one multiples the estimation from the FO protocol by ℓ. Without padding, the probability that an item is sampled is difficult to assess, making accurate frequency estimation difficult.

In [15], the FO protocol is used in a black-box fashion. That is, in order to satisfy ϵ-LDP, the FO protocol is invoked with the same privacy parameter ϵ. We observe that, since the sampling step randomly selects an item, it has an amplification effect in terms of privacy. This effect has been observed and studied in the standard DP setting [12]. If one applies an algorithm to a dataset randomly sampled from the input with a sampling rate of $\beta < 1$, to satisfy ϵ-DP, the algorithm can use a privacy budget of $\epsilon' > \epsilon$; more specifically, the relationship between ϵ', ϵ, and β is $\frac{e^{\epsilon'}-1}{e^{\epsilon}-1} = \frac{1}{\beta}$.

Intuitively, one can apply the same observation here. Since each item is selected with probability $\beta = \frac{1}{\ell}$, to satisfy ϵ-LDP, one can invoke the FO protocol with ϵ', such that $\frac{e^{\epsilon'}-1}{e^{\epsilon}-1} = \ell$ (or, equivalently $\epsilon' = \ln(\ell \cdot (e^{\epsilon} - 1) + 1) \geq \epsilon$). Surprisingly, in our study of **padding-and-sampling-based frequency oracle** (PSFO), we found that one cannot always get this privacy amplification effect. Whether this benefit is applicable or not depends on the internal structure of the FO protocol. In [18], the three best performing FO protocols are Generalized Random Response, Optimized Unary Encoding, and Optimized Local Hash. The latter two offer the same accuracy, and Optimized Local Hash has lower communication cost. It was found that Generalized Random Response offers the best accuracy when $|D| < 3e^{\epsilon} + 2$, and Optimized Local Hash offers the best accuracy when $|D| \geq 3e^{\epsilon} + 2$. We found that, the privacy amplification effect exists for Generalized Random Response, but not for Optimized Local Hash. Optimized Local Hash is able to provide better accuracy when $|D|$ is large because each perturbed output can be used to support multiple input values. However, the same feature makes Optimized Local Hash unable to benefit

from sampling. The difference in the ability to benefit from sampling changes the criterion to decide which of Generalized Random Response and Optimized Local Hash to use. We thus propose to adaptively select the best FO protocol in PSFO, based on $|I|, \epsilon$ and the particular ℓ value. Essentially, when $|I| > (4\ell^2 - \ell) \cdot e^{\epsilon} + 1$, Generalized Random Response should be used. Replacing the FO protocol used in [15] with such an adaptively chosen FO protocol greatly improves the accuracy of the resulting frequent items.

We also observe that the selection of an appropriate ℓ is crucial, and it can be different depending on the goal. Essentially, each user pads her itemset to size ℓ, generating two sources of errors: When ℓ is small, one would under-estimate the frequency counts, since items in a set with more than ℓ items will be sampled with probability less than $1/\ell$. On the other hand, since ℓ is multiplied to a noisy estimate, increasing ℓ magnifies the noises. The LDPMiner protocol in [15] has two phases, the first phase selects $2k$ candidate frequent items using a quite large ℓ, and the second phase computes their frequencies using $\ell = 2k$. We observe that for the purpose of identifying candidates for the frequent items, setting $\ell = 1$ is fine. While the resulting frequency counts under-estimate the true counts, the frequencies of all items are under-estimated, and it is very unlikely that the true top k items are not among the $2k$ candidates. However, when the goal is to estimate frequency, one needs select a larger ℓ. But ℓ should not be increased to the point that there is absolutely no under-estimation, because this increases the magnitude of noises. Selecting ℓ is a trade-off between under-estimation and noise.

Following these insights, we propose Set-Value Item Mining (SVIM) protocol, which handles set values under the LDP setting and provides much better accuracy than existing protocols within the same privacy constraints. There are four steps: First, users use PSFO with a small ℓ to report; the aggregator identifies frequent items as candidates, and sends this set to users. Second, users report (using a standard FO protocol) the number of candidate items they have; the aggregator estimates the distribution of how many candidate items the users have and selects appropriate ℓ, and sends ℓ to users. Third, users use PSFO with the given ℓ to report occurrences of items in the candidate set; the aggregator estimates the frequency of these items. Fourth, the aggregator selects the top k frequent items and use the size distribution in step two to further correct undercounts. Experimental results how that SVIM significantly outperforms LDPMiner in that it identifies more frequent items as well as estimates the frequencies more accurately.

In the setting where each user's input data is a set of items, a natural problem is to find frequent itemsets. Frequent itemset mining (FIM) is a well recognized data-mining problem. The discovery of frequent itemsets can serve valuable economic and research purposes, e.g., mining association rules [3], predicting user behavior [2], and finding correlations [7]. FIM while satisfying DP in the centralized setting has been studied extensively, e.g., [5], [19], [13]. However, because of the challenges of dealing with set-valued inputs in the LDP setting, no solution for the LDP setting has been proposed. Authors of [15] consider only the identification of frequent items, and leave FIM as an open problem. Using the PSFO technique, we are able to provide the first solution to FIM in the LDP setting. We call the protocol Set-Value itemSet Mining (SVSM) protocol; experimental evaluations demonstrates its effectiveness.

REFERENCES

[1] Mozilla governance: Usage of differential privacy & rappor. https://groups.google.com/forum/#!topic/mozilla.governance/81gMQeMEL0w.

[2] E. Adar, D. S. Weld, B. N. Bershad, and S. S. Gribble. Why we search: visualizing and predicting user behavior. In *Proceedings of the 16th international conference on World Wide Web*, pages 161–170. ACM, 2007.

[3] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[4] R. Bassily and A. Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 127–135. ACM, 2015.

[5] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta. Discovering frequent patterns in sensitive data. In *KDD*, pages 503–512, 2010.

[6] A. Bittau, U. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, and B. Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 441–459. ACM, 2017.

[7] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Acm Sigmod Record*, volume 26, pages 265–276. ACM, 1997.

[8] C. Dwork. Differential privacy. In *ICALP*, pages 1–12, 2006.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

[10] Ú. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

[11] G. Fanti, V. Pihur, and Ú. Erlingsson. Building a rappor with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies (PoPETS)*, issue 3, 2016, 2016.

[12] N. Li, W. Qardaji, and D. Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *ASIACCS*, 2012.

[13] N. Li, W. H. Qardaji, D. Su, and J. Cao. Privbasis: Frequent itemset mining with differential privacy. *VLDB*, 5(11):1340–1351, 2012.

[14] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, and J. Shin. Collecting and analyzing data from smart device users with local differential privacy. *arXiv preprint arXiv:1606.05053*, 2016.

[15] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren. Heavy hitter estimation over set-valued data with local differential privacy. In *CCS*, 2016.

[16] A. G. Thakurta, A. H. Vyrros, U. S. Vaishampayan, G. Kapoor, J. Freudiger, V. R. Sridhar, and D. Davidson. Learning new words, Mar. 14 2017. US Patent 9,594,741.

[17] A. G. Thakurta, A. H. Vyrros, U. S. Vaishampayan, G. Kapoor, J. Freudinger, V. V. Prakash, A. Legendre, and S. Duplinsky. Emoji frequency detection and deep link frequency, July 11 2017. US Patent 9,705,908.

[18] T. Wang, J. Blocki, N. Li, and S. Jha. Locally differentially private protocols for frequency estimation. In *USENIX'17: Proceedings of 26th USENIX Security Symposium on USENIX Security Symposium*. USENIX Association, 2017.

[19] C. Zeng, J. F. Naughton, and J.-Y. Cai. On differentially private frequent itemset mining. *Proceedings of the VLDB Endowment*, 6(1):25–36, 2012.