PrivBasis: Frequent Itemset Mining with Differential Privacy
by N. Li, W. Qardaji, D. Su, and J. Cao

1 Summary

In this paper authors introduced PrivBasis, a method of publishing frequent itemsets with differential privacy guarantees. Differential privacy is achieved by adding noise to frequency of itemsets, therefore participation of a single itemset in the dataset is protected. The presented method limits the search space of frequent itemsets, which speeds up the process significantly. Such subset of items, which could build any frequent itemsets, is described using basis sets, such that each frequent itemset is the subset of a basis set item. Authors introduced algorithms to privately construct basis sets, and reconstruct the frequencies of all subsets of each basis with reasonable accuracy and privacy. Having such reconstructed subsets is enough to select the most frequent itemsets.

The paper is well written, and is easy to understand. However, it has a few minor flaws, which should be solved, or at least explained broadly. For example, authors tested their approach extensively, but their approach require some a priori knowledge about dataset in order to define \( \lambda_2 \) parameter.

2 Strong Points

- Good complexity analysis, which allows to identify the parameter with exponential effect, i.e., the number \( l \) of the most frequent items in the dataset \( D \). With such analysis authors limits values of \( l \) used in their experiments \( (l < \log_3(|D|)) \) to make Algorithm 1 computable in a polynomial time.
- The paper presents extensive experiments over five datasets.
- The paper is well written and quite easy to understand. In addition, authors combined two differential privacy mechanisms – exponential and Laplace.

3 Weak Points

- Algorithm 1 is presented and some of its properties are shown and proved (achievement of differential privacy, complexity) before explaining it fully (especially lines 21–23).
- Improving accuracy of some itemsets, which can be found in two overlapping bases, is done by the weighted average with weights being a function of variance. However, this estimator of true values may be biased, which was neither analyzed nor verified.
- When considering variance in settings with multiple basis authors can reduce more than half the noise variance for the frequency of each item if \( l = 3 \). However, authors did not mentioned that a reduction is possible only for \( l < 7 \), which is a solution of \( 2^{l-1}/l^2 < 1 \) for \( l \in \mathbb{N}_+ \). There is no reduction for other values of \( l \).

4 Questions and Discussion Points

Privacy Budget Distribution. Distribution of differential privacy budget over different stages of the PrivBasis algorithm was done experimentally. It would be interesting to verify if such choice is still good enough for other datasets, and further, what properties of a dataset are important to divide the budget into optimal shares. Finding the best way of budget allocation would significantly increase the contribution of authors, but it would require additional study and a separate paper to present it.

Parameter \( \lambda_2 \). Value of the \( \lambda_2 \) parameter is defined by a heuristic formula (Section 4.4). It is important to find an exact formula for \( \lambda_2 \), which may require using some properties of a dataset. Authors do not mention how such heuristic formula was computed and do not analyze if it breaches differential privacy. More careful analysis of it is necessary.

Paper Presentations – Preferences.