CERIAS

The Center for Education and Research in Information Assurance and Security

Top-k Frequent Itemsets via Differentially Private FP-trees

Jaewoo Lee and Chris Clifton

Department of Computer Science, Purdue University

Frequent Itemset Mining

- Find all itemsets whose support is above threshold τ
- Frequent itemsets are aggregates over many individuals
- Releasing the exact result may reveal sensitive personal information

Differential Privacy

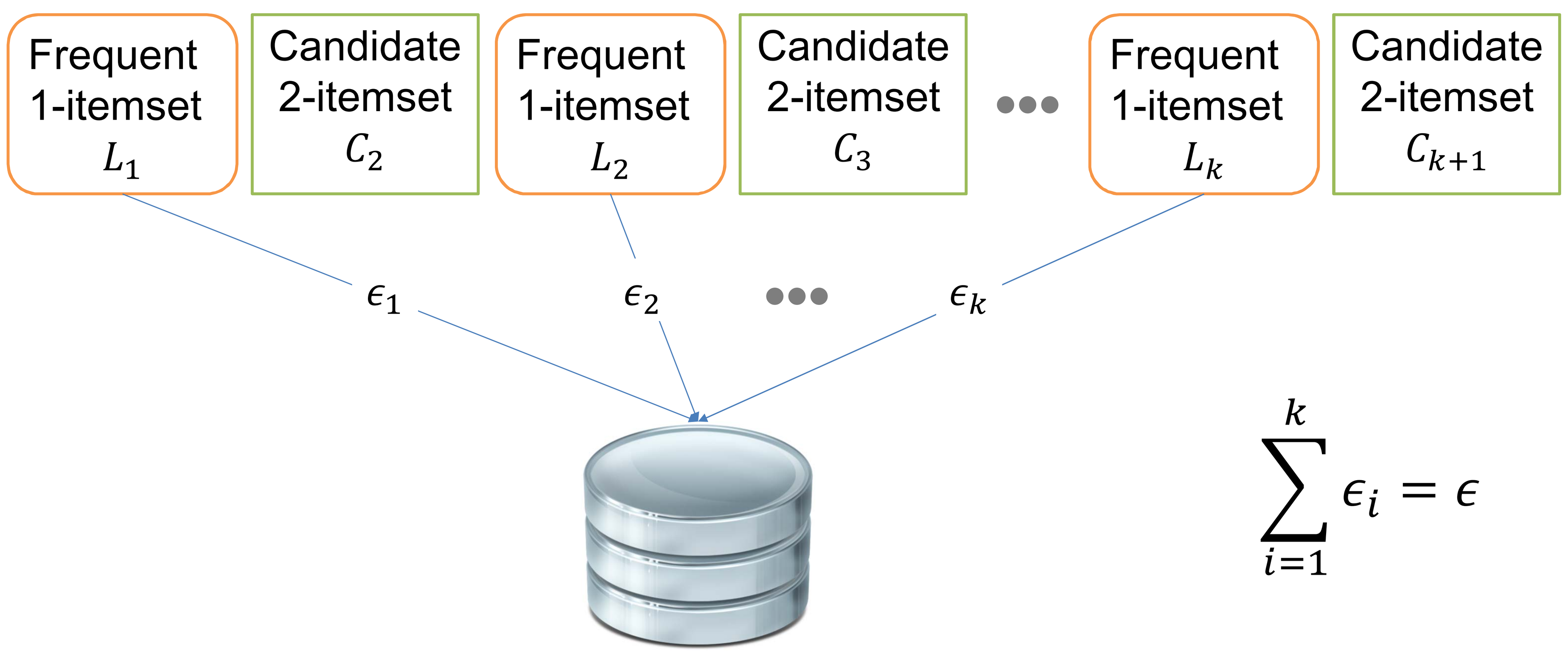
For all datasets D_1 and D_2 differing at most one element,

$$\frac{Pr[\mathcal{M}_f(D_1) = R]}{Pr[\mathcal{M}_f(D_2) = R]} \leq e^{\epsilon}$$

- output of an algorithm is insensitive to the change of a single record
- each database access costs a privacy budget

Challenge

- Given a set of items $\mathbb{I}$, the size of the search space is $O(2^{|\mathbb{I}|})$
- How to allocate privacy budget
- Smaller privacy budget implies less accurate answers
- The accuracy of algorithm is dependent on the number of queries



Our Approach

- (Phase 1) Frequent Itemset discoery
- (Phase 2) Noisy support derivation

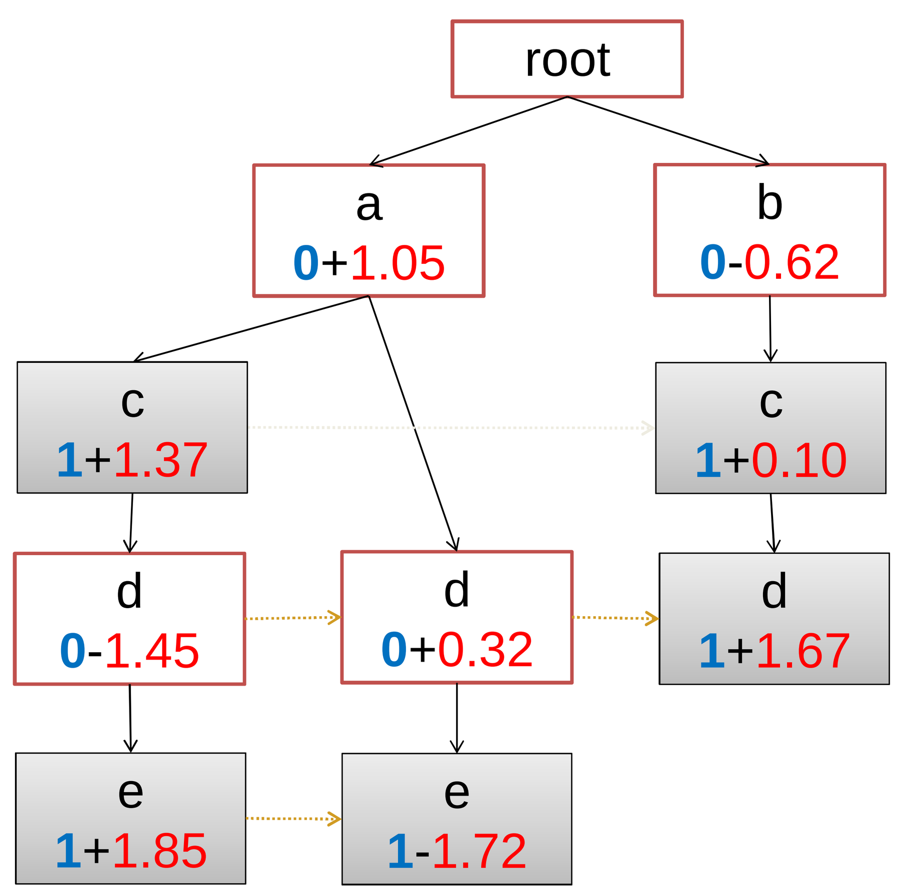
Sparse Vector Techinque

- A technique to avoid spending too much privacy budget on uninteresting queries
- Introduce a new randomness by perturbing the threshold

Algorithm 1

- $\hat{\tau} = \tau + \text{Lap}\left(\frac{2}{\epsilon}\right)$
- $\hat{X} = \sigma(X) + \text{Lap}\left(\frac{2}{\epsilon}\right)$
- If $\hat{X} \geq \hat{\tau}$ (X is frequent) then, output 1
- Otherwise (X is infrequent), output 0
- The output of algorithm is a binary vector $v = (v_1, v_2, \cdots v_t)$

Algorithm 2

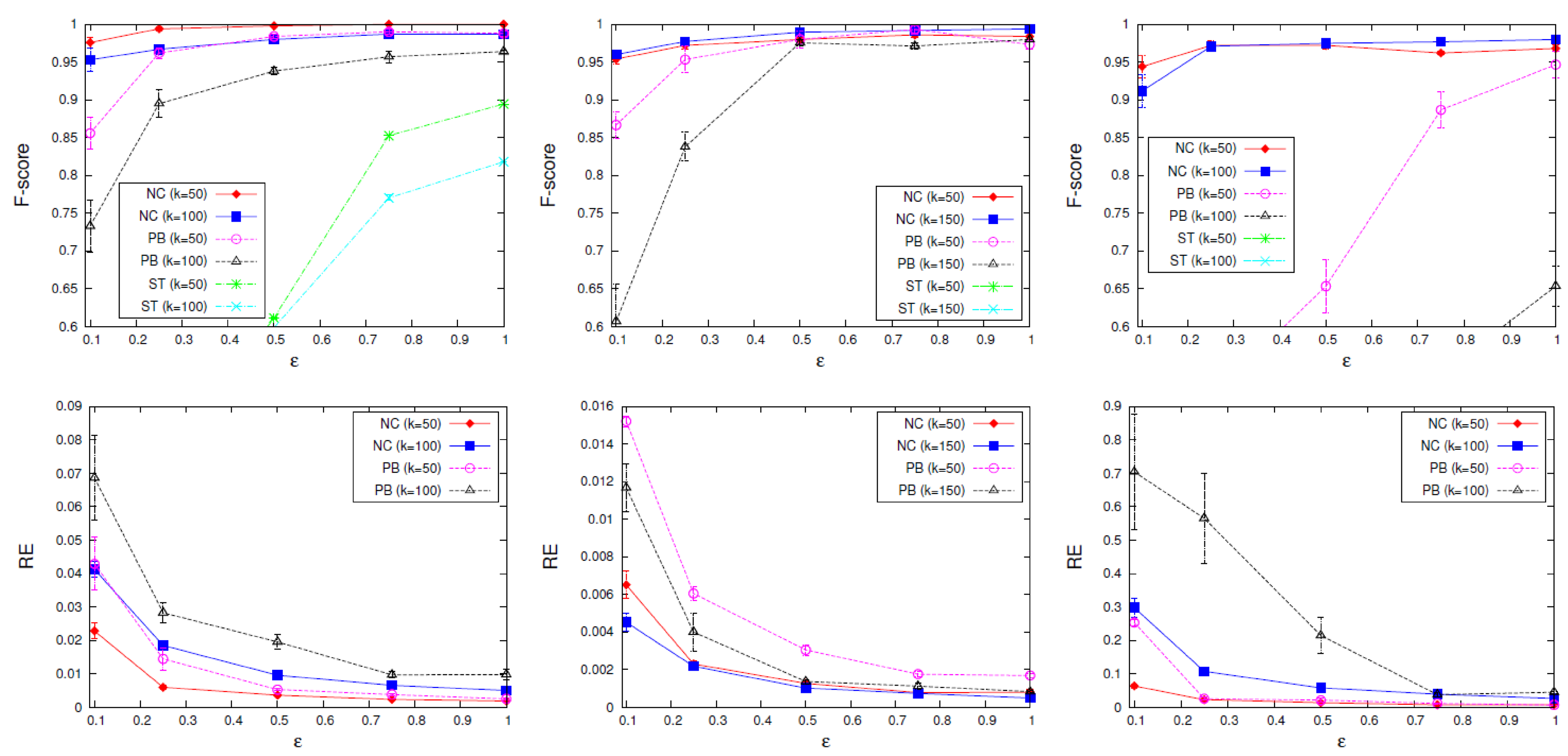


- Each node monitors the count of a prefix
- Node count is initialized with a noise
- To get the correct count, child's count needs to be added to its parent's count
- (optional) post-processing can increase the accuracy

Performance Evaluation

- F-score = $\frac{2(\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}$
- Relative error = $\text{median}_X \left(\frac{|\hat{\sigma}(X) - \sigma(X)|}{\sigma(X)} \right)$
- the proposed method outperforms other two methods throughout all test datasets

dataset	$\|D\|$	$\|\mathcal{I}\|$	max $\|t\|$	avg $\|t\|$
mushroom (MUS)	8,124	119	23	23
pumsb star (PUMSB)	49,046	2,088	63	50.5
retail (RETL)	88,162	16,470	76	10.3



(a) mushroom (b) pumsb star (c) retail

PURDUE

UNIVERSITY