Supplement Experiment Results of ϵ-PPI

Yuzhe Tang [†] Ling Liu [‡]

[†]*Georgia Institute of Technology, GA, USA, Email:* yztang@gatech.edu
[‡]*Georgia Institute of Technology, GA, USA, Email:* lingliu@cc.gatech.edu

I. Experiments

A. Privacy on Multi-term Indexing

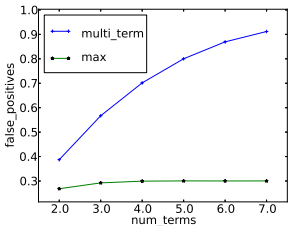

Fig. 1: Multi-term false positive rates with varying number of terms

We have performed experiments to verify the property of ϵ-PPI to protect privacy of multi-term phrases. Before the experiment, we prepare a synthetic dataset by generating term-incidence matrix following pre-defined frequencies. In specific, data is generated for x terms (where $x \in [2, 7]$) and each term has certain frequency. We write our data generator in such way that terms are independently distributed and there is at least one co-appearance of all terms. In the first experiment, we test the false positive rates against different number of terms. We vary the number of terms from 2 to 7, but for each term, fix its ϵ to be the same value. We use the data file generated with term frequency fixed at 0.25 for all terms. Two metrics are used, that is, false positive rate for multi-term phrase (denoted by $multi\_term$) and maximal single-term false positive rate among all the terms (denoted by max). Experiments have been repeated 1000 times and the average results are shown in Fig. 1, in which it can be seen that the false positive rate for multi-term phrases is always higher than that for the maximal single-term phrases. In particular, as number of terms grows in a phrase, the discrepancy increases. The experimental result, in addition to our previous analysis result, shows the even higher level of privacy preservation for multi-term phrases than of an individual term.

TABLE I: Multi-term false positive rates with varying distributions

	Frequency distr.	ϵ distr.	multi_term	max
1	0.25:0.25:0.25	0.3:0.3:0.3	0.5652	0.2948
2	0.25:0.25:0.25	0.1:0.2:0.3	0.3972	0.2941
3	0.25:0.25:0.25	0.1:0.3:0.5	0.6182	0.4974
4	0.125:0.25:0.5	0.3:0.3:0.3	0.5789	0.2975
5	0.125:0.25:0.5	0.1:0.2:0.3	0.3754	0.2967
6	0.125:0.25:0.5	0.3:0.2:0.1	0.3344	0.2759

To further study the privacy preservation, we test with different distributions in term frequency and sensitivity ϵ. With the results shown in Table I, we starts with a baseline uniform configuration (as in line 1), in which frequency and ϵ for all

terms are fixed at 0.25 and 0.3, respectively. With ϵ changed to be non-uniform or skewed (e.g., 0.1, 0.2 and 0.3 as shown in line 2), multi-term false positives decreases but still bigger than the maximal single-term false positive. With ϵ changed to $0.1, 0.3, 0.5$ as in line 3, the maximal false positive rate increases but still smaller than the multi-term one. As demonstrated by line $x4, 5, 6$, the multi-term false positive rate is not sensitive to term frequency distribution.