Collecting Individual Trajectories under Local Differential Privacy (Technical Report)

Jianyu Yang[1], Xiang Cheng[2*], Sen Su[3], Huizhong Sun[4], Changju Chen[5]
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing, China
{jyyang[1], chenchangjv[5]}@bupt.cn {chengxiang[2], susen[3], showlostage1[4]}@bupt.edu.cn

I. Supplementary Analysis

A. Detailed Derivation of Squared Sampling and Noise Error

In our scenario, we suppose the n users of a dataset D are randomly divided into t groups. We analyse the estimation run on a subdataset D_φ, i.e., one of the t groups. We use $f_v(D)$ and $\bar{f}_v(D)$ to denote the estimated and true frequencies of the value v in the dataset D, respectively. For simplicity, the frequency on the original dataset $\bar{f}_v(D)$ is written as $\bar{f}_v$. The expected squared sampling and noise error for estimating one value is

$$\begin{aligned}
&\mathbf{E}\left[\left(f_v(D_\varphi)-\bar{f}_v\right)^2\right]\\
=&\mathbf{E}\left[\left((f_v(D_\varphi)-\bar{f}_v(D_\varphi))+(\bar{f}_v(D_\varphi)-\bar{f}_v)\right)^2\right]\\
=&\mathbf{E}\left[\left(f_v(D_\varphi)-\bar{f}_v(D_\varphi)\right)^2\right]+\mathbf{E}\left[\left(\bar{f}_v(D_\varphi)-\bar{f}_v\right)^2\right]+\\
&2\mathbf{E}\left[(f_v(D_\varphi)-\bar{f}_v(D_\varphi))\cdot(\bar{f}_v(D_\varphi)-\bar{f}_v)\right] \qquad (1)
\end{aligned}$$

Specifically, Equation (1) consists of three parts. The first part is the variance of OLH, i.e.,

$$\begin{aligned}
&\mathbf{E}\left[\left(f_v(D_\varphi)-\bar{f}_v(D_\varphi)\right)^2\right]\\
=&t\cdot\frac{p^*(1-p^*)+\bar{f}_v(p-p^*)(1-p-p^*)}{n(p-p^*)^2}\\
=&t\cdot\frac{p^*(1-p^*)}{n(p-p^*)^2}+t\cdot\frac{\bar{f}_v(p-p^*)(1-p-p^*)}{n(p-p^*)^2}\\
=&\frac{4te^\varepsilon}{n(e^\varepsilon-1)^2}+\frac{t}{n}\cdot\bar{f}_v \quad \text{(for OLH, } p=1/2 \text{ and } p^*=\frac{1}{e^\varepsilon+1}\text{)}.
\end{aligned}$$

*Corresponding author

The second part is

$$\begin{aligned}
&\mathbf{E}\left[\left(\bar{f}_v(D_\varphi)-\bar{f}_v\right)^2\right]\\
=&\mathbf{E}\left[\bar{f}_v^2(D_\varphi)\right]-2\bar{f}_v\mathbf{E}\left[\bar{f}_v(D_\varphi)\right]+\bar{f}_v^2\\
=&\mathbf{E}\left[\bar{f}_v^2(D_\varphi)\right]-\bar{f}_v^2\\
=&\mathbf{E}\left[\left(\frac{t}{n}\sum\mathbb{1}_{\{v_i=v\}}\right)^2\right]-\bar{f}_v^2\\
=&\left(\frac{t}{n}\right)^2\mathbf{E}\left[\left(\sum\mathbb{1}_{\{v_i=v\}}\right)^2\right]-\bar{f}_v^2\\
=&\left(\frac{t}{n}\right)^2\mathbf{E}\left[\sum_i\mathbb{1}^2_{\{v_i=v\}}+\sum_{i\neq j}\mathbb{1}_{\{v_i=v\}}\cdot\mathbb{1}_{\{v_j=v\}}\right]-\bar{f}_v^2\\
=&\left(\frac{t}{n}\right)^2\left[\frac{n}{t}\bar{f}_v+\left(\frac{n^2}{t^2}-\frac{n}{t}\right)\bar{f}_v\cdot\frac{n\bar{f}_v-1}{n-1}\right]-\bar{f}_v^2\\
=&\frac{t}{n}\bar{f}_v+\left(1-\frac{t}{n}\right)\bar{f}_v\cdot\frac{n\bar{f}_v-1}{n-1}-\bar{f}_v^2\\
=&\left(\frac{t}{n}-\frac{n-t}{n}\frac{1}{n-1}\right)\bar{f}_v+\left(1-\frac{t}{n}\right)\bar{f}_v\cdot\frac{n\bar{f}_v}{n-1}-\bar{f}_v^2\\
=&\frac{t-1}{n-1}\bar{f}_v(1-\bar{f}_v).
\end{aligned}$$

The third part is

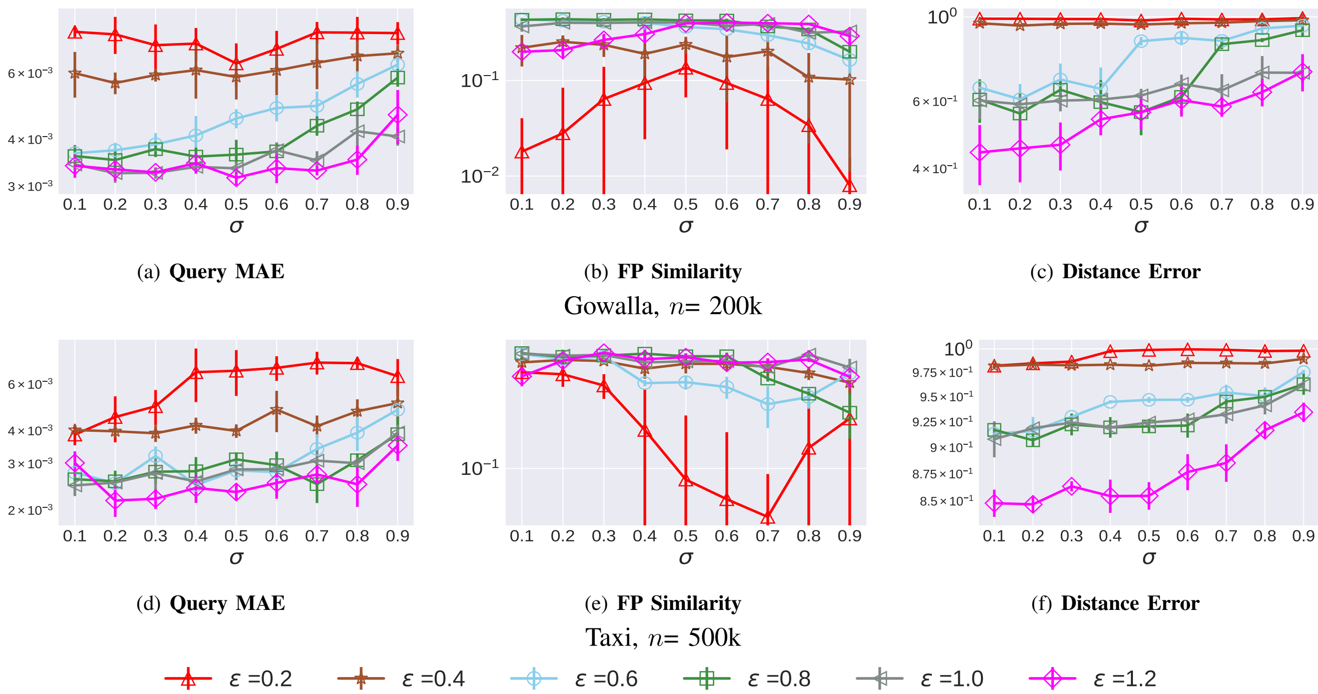
$$\begin{aligned}
&2\mathbf{E}\left[(f_v(D_\varphi)-\bar{f}_v(D_\varphi))\cdot(\bar{f}_v(D_\varphi)-\bar{f}_v)\right]\\
=&2\mathbf{E}\left[(f_v(D_\varphi)-\bar{f}_v(D_\varphi))\cdot\bar{f}_v(D_\varphi)\right]\\
&\text{(as } \mathbf{E}[f_v(D_s)]=\mathbf{E}[\bar{f}_v(D_s)] \text{ and } \bar{f}_v \text{ is a constant)}\\
=&2\mathbf{E}\left[\mathbf{E}\left[(f_v(D_\varphi)-\bar{f}_v(D_\varphi))\cdot\bar{f}_v(D_\varphi)\mid D_\varphi\right]\right]\\
=&0.
\end{aligned}$$

We observe that the second part is a constant which is much smaller than the first part. Ignoring the small factor $\frac{t}{n}\cdot\bar{f}_v$ in the first part, the expected squared sampling and noise error can be dominated by $\frac{4te^\varepsilon}{n\cdot(e^\varepsilon-1)^2}$.

B. Effectiveness of Guideline

To further judge the effectiveness of our guideline for choosing granularities in PrivAG, we validate the recommended settings of parameters σ and α.

Impact of σ. In PrivTC, the n users are divided into two groups U_1 and U_2, where U_1 has a population of $|U_1|=n\cdot\sigma$



(a) **Query MAE** (b) **FP Similarity** (c) **Distance Error**

Gowalla, n= 200k

(d) **Query MAE** (e) **FP Similarity** (f) **Distance Error**

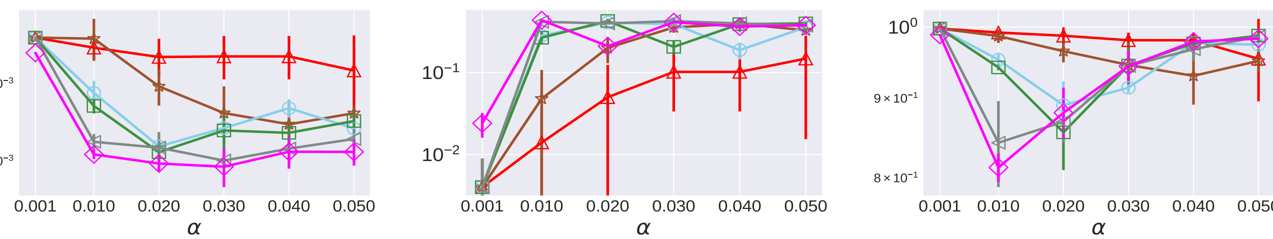
Taxi, n= 500k

Fig. 1: **PrivTC varying σ under setting of t= 9. Results are shown in log scale.**

while U_2 has $|U_2| = n \cdot (1 - \sigma)$. In particular, U_1 is used to adaptively partition the 2-D domain into a grid in PrivAG, and U_2 is used for learning the HMM in PrivSL.

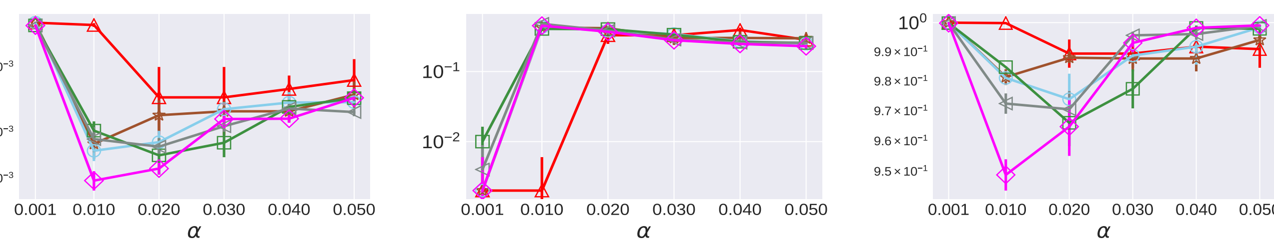
Figure 1 shows the results of PrivTC with different ε values varying σ from 0.1 to 0.9. From Figure 1, we can see that in all cases, the values of σ ranging from 0.1 to 0.3 can make PrivTC typically achieve nearly best the best performance, which confirms the effectiveness of our recommended setting of $\sigma = 0.2$. The intuition behind this is that a relatively small number of users in group U_1 is sufficient for PrivAG to construct a reasonable grid. Assigning a larger population to the group U_2 may help learn the accurate parameters of HMM in PrivSL, which plays a more important role for boosting the final utility.

Impact of α. Figure 2 studies the impact of α on the utility of PrivTC with different ε values varying α from 0.001 to 0.05. We can observe that setting α in the range of $[0.01, 0.02]$ can typically obtain the good performance of PrivTC, which verify the effectiveness of our recommended setting of $\alpha = 0.02$. In particular, the utility of PrivTC has a huge improvement when α changes from 0.001 to 0.01 and usually degrades as α is larger than 0.02. The reason is that as a constant related to grid construction, too small values of α such as 0.001 will lead to insufficient partitioning of the 2-D spatial domain, losing the statistic features of the original trajectories; while relatively larger values may over partition the spatial domain, resulting in too excessive noise errors.



(a) **Query MAE** (b) **FP Similarity** (c) **Distance Error**

Gowalla, n= 200k



(d) **Query MAE** (e) **FP Similarity** (f) **Distance Error**

Taxi, n= 500k

ε =0.2 ε =0.4 ε =0.6 ε =0.8 ε =1.0 ε =1.2

Fig. 2: **PrivTC varying α under setting of t= 9. Results are shown in log scale.**