Enhanced Accuracy in Local Differential Privacy via Leveraging Inherent Uncertainty (Technical Report)

Peng Tang [†1], Xiya Shao [†2], Rui Chen [‡], Ning Wang [♮], Shanqing Guo [†3]
[†]School of Cyber Science and Technology, Shandong University, China
[†]{tangpeng[1], guoshanqing[3]}@sdu.edu.cn, shaoxiya[2]@mail.sdu.edu.cn
[‡]College of Computer Science and Technology, Harbin Engineering University, China, ruichen@hrbeu.edu.cn
[♮] Cyberspace Institute of Advanced Technology, Guangzhou University, China, wangning@gzhu.edu.cn

I. Proofs

A. Proof of Lemma 1

Proof 1: When $\frac{\alpha}{\beta} > e^{\varepsilon}$, according Lemma 1, we know that

$$\frac{\alpha p(1-q)+(1-\alpha)(1-p)q}{\beta p(1-q)+(1-\beta)(1-p)q} \leq e^{\varepsilon}. \tag{1}$$

To maximize data utility, let the equality hold, i.e.,

$$\frac{\alpha p(1-q)+(1-\alpha)(1-p)q}{\beta p(1-q)+(1-\beta)(1-p)q} = e^{\varepsilon}. \tag{2}$$

It can be derived that,

$$p = g(q) = \frac{(1-e^{\varepsilon})q-(\alpha-\beta e^{\varepsilon})q}{(1-e^{\varepsilon})q-(\alpha-\beta e^{\varepsilon})}. \tag{3}$$

We denote $\mu = 1-e^{\varepsilon}$, $\omega = \alpha-\beta e^{\varepsilon}$, then,

$$u(q) = \mu q - \omega q, \quad v(q) = \mu q - \omega. \tag{4}$$

The variance of the estimated value obtained by the AOUE protocol is:

$$\text{var} = \frac{nq(1-q)}{(p-q)^2} = \frac{nq(1-q)}{(g(q)-q)^2}.$$

To minimize the variance, we take the derivative of the variance with respect to q and set it to zero.

$$\begin{aligned}\frac{\partial \text{var}}{\partial q} =& n(1-q)(g(q)-q)^{-2} - nq(g(q)-q)^{-2} \\ & - 2nq(1-q)(g(q)-q)^{-3}(g'(q)-1)\end{aligned}.$$

Where,

$$g'(q) = \frac{u'(q)v(q)-u(q)v'(q)}{v^2(q)},$$

$$u'(q) = (1-e^{\varepsilon}) - (\alpha-\beta e^{\varepsilon}) = \mu - \omega,$$

$$v'(q) = 1-e^{\varepsilon} = \mu.$$

Then it follows that,

$$(1-2q)(g(q)-q) - 2q(1-q)(g'(q)-1) = 0.$$

Further deriving,

$$(1-2q)\left(\frac{u(q)}{v(q)}-q\right) - 2q(1-q)\left(\frac{u'(q)v(q)-u(q)v'(q)}{v^2(q)}-1\right) = 0.$$

Continuing to derive the above equation,

$$\begin{aligned}&(1-2q)(u(q)v(q)-qv^2(q)) \\ &-2q(1-q)(u'(q)v(q)-u(q)v'(q)-v^2(q)) \\ =&(1-2q)u(q)v(q) - q(1-2q)v^2(q) \\ &+2q(1-q)v^2(q) - 2(1-q)(qu'(q)v(q)-qu(q)v'(q)) \\ =&(1-2q)u(q)v(q) + qv^2(q) - 2(1-q)(u(q)v(q)-qu(q)v'(q)) \\ =&-u(q)v(q)+qv^2(q)+2(1-q)qu(q)v'(q) \\ =&-(\mu q-\omega q)(\mu q-\omega)+q(\mu q-\omega)^2+2(1-q)q(\mu q-\omega q)\mu \\ =&(-\mu^2q^2+\mu\omega q+\mu\omega q^2-\omega^2 q)+(\mu^2q^3-2\mu\omega q^2+\omega^2 q) \\ &+(2\mu^2q^2-2\mu\omega q^2-2\mu^2q^3+2\mu\omega q^3) \\ =&(-\mu^2+2\mu\omega)q^3+(\mu^2-3\mu\omega)q^2+\mu\omega q \\ =&-\mu q((\mu-2\omega)q^2-(\mu-3\omega)q-\omega) \\ =&-\mu q((\mu-2\omega)q+\omega)(q-1) = 0.\end{aligned}$$

Therefore,

$$q = \frac{\omega}{2\omega-\mu} = \frac{\alpha-\beta e^{\varepsilon}}{(e^{\varepsilon}-1)+2(\alpha-\beta e^{\varepsilon})}, \tag{5}$$

$$p = \frac{\mu q-\omega q}{\mu q-\omega} = \frac{\omega(\mu-\omega)}{\mu\omega-2\omega^2+\mu\omega} = \frac{1}{2}. \tag{6}$$

B. Proof of Theorem 4

Proof 2: To prove that the two-stage scheme satisfies ε-LDP, we need to prove that both 10 and 11 in Problem 2 hold.

We first prove that Eq. 10 holds. We denote the missing state as m_1, where $m_1 \in M$. Its perturbed state is denoted as $\widetilde{m}_1 \in \widetilde{M}$. W.l.o.g., let

$$m_1 = \arg\max_{m\in\Omega_M}\left\{\frac{\max_{s_i\in\Omega_S}\{\Pr(M=m|S=s_i)\}}{\min_{s_j\in\Omega_S}\{\Pr(M=m|S=s_j)\}}\right\}.$$

For any $s_i \in S$, we can calculate the conditional probability

as

$$\begin{aligned}
&\Pr(\widetilde{M}=\widetilde{m}_1|S=s_i)\\
=&\Pr(M=m_1|S=s_i)\cdot\Pr(\widetilde{M}=\widetilde{m}_1|M=m_1)\\
&+\Pr(M=m_2|S=s_i)\cdot\Pr(\widetilde{M}=\widetilde{m}_1|M=m_2)\\
=&\Pr(M=m_1|S=s_i)\cdot p_2+\Pr(M=m_2|S=s_i)\cdot q_2\\
=&\Pr(M=m_1|S=s_i)\cdot p_2+(1-\Pr(M=m_1|S=s_i))\cdot q_2.
\end{aligned}$$

Since $p_2 > q_2$ and $\beta \le \Pr(M=m_1|S=s_i) \le \alpha$, we have

$$\beta\cdot p_2+(1-\beta)\cdot q_2 \le \Pr(\widetilde{M}=\widetilde{m}_1|S=s_i) \le \alpha\cdot p_2+(1-\alpha)\cdot q_2.$$

Therefore, for $\forall s_i, s_j \in S$,

$$\frac{\Pr(\widetilde{M}=\widetilde{m}_1|S=s_i)}{\Pr(\widetilde{M}=\widetilde{m}_1|S=s_j)} \le \frac{\alpha\cdot p_2+(1-\alpha)\cdot q_2}{\beta\cdot p_2+(1-\beta)\cdot q_2} \le e^{\varepsilon}.$$

Moreover, we can prove that Eq. 10 still holds when the output value is a non-missing value $\widetilde{m}_2$. Since $\Pr(\widetilde{M}=\widetilde{m}_2|S=s_i)=\sum_{k=1}^{d}\Pr(\widetilde{A}=\widetilde{a}_k|S=s_i)$, we first calculate $\Pr(\widetilde{A}=\widetilde{a}_1|S=s_i)$.

$$\begin{aligned}
&\Pr(\widetilde{A}=\widetilde{a}_1|S=s_i)\\
=&\Pr(M=m_1|S=s_i)\cdot\Pr(\widetilde{A}=\widetilde{a}_1|M=m_1)\\
&+\Pr(M=m_2|S=s_i)\cdot\Pr(\widetilde{A}=\widetilde{a}_1|M=m_2)\\
=&\Pr(M=m_1|S=s_i)\cdot q_2\\
&+\Pr(M=m_2|S=s_i)\cdot\Pr(\widetilde{A}=\widetilde{a}_1|M=m_2).
\end{aligned}$$

In the above equation,

$$\begin{aligned}
&\Pr(\widetilde{A}=\widetilde{a}_1|M=m_2)=\frac{\Pr(\widetilde{A}=\widetilde{a}_1,M=m_2)}{\Pr(M=m_2)}\\
=&\sum_{k=1}^{d}\frac{\Pr(A=a_k)\Pr(\widetilde{A}=\widetilde{a}_1|A=a_k)}{\sum_{k'=1}^{d}\Pr(A=a_{k'})}\\
\ge&\min_k\{\Pr(\widetilde{A}=\widetilde{a}_1|A=a_k)\}\\
=&\Pr(\widetilde{A}=\widetilde{a}_1|A=a_{k\neq 1})\\
=&q^* > q_2.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&(1-\alpha)\cdot\Pr(\widetilde{A}=\widetilde{a}_1|M=m_2)+\alpha\cdot q_2\\
\le&\Pr(\widetilde{A}=\widetilde{a}_1|S=s_i)\\
\le&(1-\beta)\cdot\Pr(\widetilde{A}=\widetilde{a}_1|M=m_2)+\beta\cdot q_2.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\Pr(\widetilde{A}=\widetilde{a}_1|M=m_2)=\frac{\Pr(\widetilde{A}=\widetilde{a}_1,M=m_2)}{\Pr(M=m_2)}\\
=&\sum_{k=1}^{d}\frac{\Pr(A=a_k)\Pr(\widetilde{A}=\widetilde{a}_1|A=a_k)}{\sum_{k'=1}^{d}\Pr(A=a_{k'})}\\
\le&\max_k\{\Pr(\widetilde{A}=\widetilde{a}_1|A=a_k)\}=\Pr(\widetilde{A}=\widetilde{a}_1|A=a_1)\\
=&p^* \le p_2.
\end{aligned}$$

Let $\Delta = p_2 - \Pr(\widetilde{A}=\widetilde{a}_1|M=m_2)$. Therefore,

$$\begin{aligned}
&(1-\alpha)\cdot(p_2-\Delta)+\alpha\cdot q_2\\
\le&\Pr(\widetilde{A}=\widetilde{a}_1|S=s_i)\\
\le&(1-\beta)\cdot(p_2-\Delta)+\beta\cdot q_2.
\end{aligned}$$

Recall that $\frac{1-\beta}{1-\alpha} \le \frac{\alpha}{\beta}$. When $\frac{1-\beta}{1-\alpha} > e^{\varepsilon}$,

$$\begin{aligned}
&\frac{(1-\beta)\cdot p_2+\beta\cdot q_2}{(1-\alpha)\cdot p_2+\alpha\cdot q_2}-\frac{\alpha\cdot p_2+(1-\alpha)\cdot q_2}{\beta\cdot p_2+(1-\beta)\cdot q_2}\\
=&\frac{(\beta(1-\beta)-\alpha(1-\alpha))\cdot(p_2^2+q_2^2-2p_2q_2)}{((1-\alpha)\cdot p_2+\alpha\cdot q_2)\cdot(\beta\cdot p_2+(1-\beta)\cdot q_2)}\\
\le&0.
\end{aligned}$$

Therefore,

$$\frac{(1-\beta)\cdot p_2+\beta\cdot q_2}{(1-\alpha)\cdot p_2+\alpha\cdot q_2} \le \frac{\alpha\cdot p_2+(1-\alpha)\cdot q_2}{\beta\cdot p_2+(1-\beta)\cdot q_2} \le e^{\varepsilon}.$$

Furthermore, since

$$(1-\beta)\cdot p_2+\beta\cdot q_2 \le e^{\varepsilon}((1-\alpha)\cdot p_2+\alpha\cdot q_2),$$

$$\text{and}\quad -\Delta(1-\beta) < e^{\varepsilon}(-\Delta(1-\alpha)),$$

we have

$$(1-\beta)\cdot(p_2-\Delta)+\beta\cdot q_2 \le e^{\varepsilon}((1-\alpha)\cdot(p_2-\Delta)+\alpha\cdot q_2).$$

Then, for $\forall s_i, s_j \in S$,

$$\frac{\Pr(\widetilde{A}=\widetilde{a}_1|S=s_i)}{\Pr(\widetilde{A}=\widetilde{a}_1|S=s_j)} \le \frac{(1-\beta)\cdot(p_2-\Delta)+\beta\cdot q_2}{(1-\alpha)\cdot(p_2-\Delta)+\alpha\cdot q_2} \le e^{\varepsilon}\quad .$$

In addition, when $\frac{1-\beta}{1-\alpha} \le e^{\varepsilon}$, obviously,

$$\frac{\Pr(\widetilde{A}=\widetilde{a}_1|S=s_i)}{\Pr(\widetilde{A}=\widetilde{a}_1|S=s_j)} \le e^{\varepsilon}.$$

Therefore,

$$\frac{\Pr(\widetilde{M}=\widetilde{m}_2|S=s_i)}{\Pr(\widetilde{M}=\widetilde{m}_2|S=s_j)}=\frac{\sum_{k=1}^{d}\Pr(\widetilde{A}=\widetilde{a}_k|S=s_i)}{\sum_{k=1}^{d}\Pr(\widetilde{A}=\widetilde{a}_k|S=s_j)} \le e^{\varepsilon}.$$

Up to this point, we have proved that Eq. 10 holds for any $\widetilde{M}$ ($\widetilde{m}_1$ or $\widetilde{m}_2$).

Then, we prove that Eq. 11 holds. For Eq. 11, it is easy to prove that

$$\frac{\Pr(\widetilde{A}=\widetilde{a}_k|A=a_i)}{\Pr(\widetilde{A}=\widetilde{a}_k|A=a_j)} \le \frac{p^*}{q^*}=e^{\varepsilon}.$$

Therefore, the two-stage scheme satisfies ε-LDP.

II. Utility Analysis of AGRR

To measure the data utility of the AGRR protocol, we compare it with the data utility of the GRR protocol. In the traditional GRR protocol, the perturbation parameter satisfies

$$\Pr\left(\widetilde{A}=\widetilde{a}|A=a\right)=\begin{cases}p'=\frac{e^{\varepsilon}}{e^{\varepsilon}+d-1}, & \text{if } \widetilde{a}=a\\ q'=\frac{1-p'}{d-1}=\frac{1}{e^{\varepsilon}+d-1}, & \text{if } \widetilde{a}\neq a\end{cases}.$$

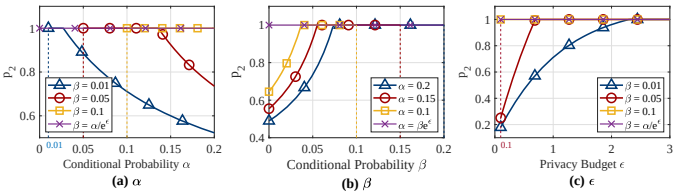

Fig. 1. The value of p_2 under different parameters.

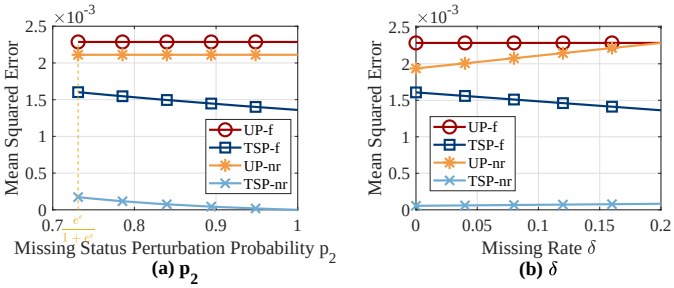

Fig. 2. MSE of TSP/UP under p_2 and δ.

We can calculate that the variance of the estimated value in the GRR protocol is:

$$\text{var}_{\text{grr}} = \frac{nq'(1-q')}{(p'-q')^2}.$$

In the AGRR protocol, the variance of the estimated value is:

$$\text{var}_{\text{agrr}} = \frac{nq(1-q)}{(p-q)^2}.$$

Since $p > \frac{e^\varepsilon}{e^\varepsilon+d} = p'$ and $p+(d-1)q = 1, p'+(d-1)q' = 1$. Therefore, $q < q'$. Thus, it follows that

$$\text{var}_{\text{agrr}} = \frac{nq(1-q)}{(p-q)^2} < \frac{nq'(1-q')}{(p'-q')^2} = \text{var}_{\text{grr}}.$$

Therefore, the data utility of the AGRR protocol is superior to the data utility of the GRR protocol.

III. Some Supplementary Experiments

We demonstrate the effectiveness of *TSP* and *UP* within a broader range of parameter values.

Fig. 1(a) shows the relationship between the probability of the missing status remaining unchanged, denoted as p_2, and the conditional probability α. We set ε to 1 and consider β values from $\{0.01, 0.05, 0.1\}$. The range of α is defined as $[\beta, 0.2]$. It can be observed that, with fixed β and ε, as α increases starting from $\alpha = \beta e^\varepsilon$ (i.e., $\frac{\alpha}{\beta} = e^\varepsilon$), p_2 continuously decreases. This implies that the perturbation scale of the missing status increases. This is because that when the *global uncertainty* requirement remains constant, as the *inherent uncertainty* of the data decreases ($\frac{\alpha}{\beta}$ gradually increases), the uncertainty introduced by LDP grows.

Fig. 1(b) shows the relationship between p_2 and β. We set $\varepsilon = 1$, and consider α values from $\{0.1, 0.15, 0.2\}$. The range of β is defined as $(0, \alpha]$. It can be observed that, given α and ε, as β increases, the probability p_1 continuously increases until $\beta = \frac{\alpha}{e^\varepsilon}$, (i.e., $\frac{\alpha}{\beta} = e^\varepsilon$). When $\beta \geq \frac{\alpha}{e^\varepsilon}$, p_2 equals 1, indicating that the missing status is directly reported without perturbation.

Fig. 1(c) shows the relationship between p_2 and ε. We set $\alpha = 0.1$, and consider β values from $\{0.01, 0.05, 0.1\}$. The range of ε is defined as $[0.1, 3]$. It can be observed that, with increasing ε, p_2 increases. The reason lies in that, when the *global uncertainty* decreases, the *perturbation uncertainty* also decreases.

Fig. 2(a) shows the performance (including the estimation of frequency and null value rate) of TSP and UP under different probability p_2. We set $\varepsilon = 1$, $\delta = 0.1$, and $d = 4$. It can be found that, our *TSP* scheme consistently achieves a lower MSE compared to UP. The MSE of UP remains unaffected by changes in p_2. In contrast, our proposed TSP scheme exhibits a decreasing trend in MSE as p_2 increases, indicating a reduction in the perturbation scale for the missing status.

Fig. 2(b) shows the performance (including the estimation of frequency and null value rate) of TSP and UP under different missing rate δ. We set $\varepsilon = 1$, $p_2 = 0.85$, and $d = 4$. It can be observed that, *TSP* consistently achieves a lower MSE compared to UP. As the missing rate increases, our solution's MSE continues to decrease. This can be attributed to the fact that, as the missing rate rises, the error of the estimation of missing value increases, while the squared error of the estimation of valid values decreases. In contrast, *TSP* exhibits a larger perturbation magnitude in valid values, resulting in a more significant change in the error of their estimation.