Mathematical Biosciences and Engineering

http://www.aimspresScom/journal/MBE

MBE, 18(3): 2733–2763.
DOI: 10.3934/mbe.2021139
Received: 01 February 2021
Accepted: 09 March 2021
Published: 22 March 2021

Research article

Set-valued data collection with local differential privacy based on category hierarchy

Jia Ouyang[1], Yinyin Xiao[1,*], Shaopeng Liu[2], Zhenghong Xiao[2] and Xiuxiu Liao[2]

[1] School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou 510665, China
[2] College of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China

* **Correspondence:** Email: gsxyy@gpnu.edu.cn; Tel: +8613430337590.

Supplementary

Theorem 4.1. Category perturbation algorithm (CP) satisfies the ε_1-local differential privacy.
Proof:

For two different inputs b_1 and b_2, it is necessary to prove that the upper bound of the ratio of the probability of the same result b^* is e^{ε_1}, and the $b^*=1$ case is proved first; $b^*=0$ can be proved by the same procedure.

$$\frac{\Pr\left[b^*\middle|b_1\right]}{\Pr\left[b^*\middle|b_2\right]}=\frac{\Pr\left[1\middle|b_1\right]}{\Pr\left[1\middle|b_2\right]}\le\frac{\Pr\left[1\middle|1\right]}{\Pr\left[1\middle|0\right]}=\frac{\frac{e^{\varepsilon_1}}{1+e^{\varepsilon_1}}}{\frac{1}{1+e^{\varepsilon_1}}}=e^{\varepsilon_1} \tag{S1}$$

Theorem 4-1 is proved.
Theorem 4.2. The Value Perturbation algorithm I (VP_LP) satisfies ε_2-LDP
Proof:

The random variable that obeys the Laplace distribution is defined as follows:

$$\Pr\left[Lap(b)=x\right]=\frac{1}{2\cdot b}\cdot\exp\left(-\frac{|x|}{b}\right) \tag{S2}$$

Given two different inputs v_1, v_2, the probability of output v' is:

$$\Pr\left[v'\middle|v_1\right]=\frac{\varepsilon_2}{2\cdot\Delta_{LP}}\cdot\exp\left(-\frac{\varepsilon_2\cdot\left|v'-v_1\right|_1}{\Delta_{LP}}\right) \tag{S3}$$

Hence,

$$\frac{\Pr\left[v'\middle|v_1\right]}{\Pr\left[v'\middle|v_2\right]}=\frac{\frac{\varepsilon_2}{2\cdot\Delta_{LP}}\cdot\exp\left(-\frac{\varepsilon_2\cdot\left|v'-v_1\right|_1}{\Delta_{LP}}\right)}{\frac{\varepsilon_2}{2\cdot\Delta_{LP}}\cdot\exp\left(-\frac{\varepsilon_2\cdot\left|v'-v_2\right|_1}{\Delta_{LP}}\right)}=\exp\left(\frac{\varepsilon_2\cdot\left(\left|v'-v_1\right|_1-\left|v'-v_2\right|_1\right)}{\Delta_{LP}}\right) \tag{S4}$$

We use the triangle inequality of the absolute value to get:

$$\frac{\Pr\left[v'\middle|v_1\right]}{\Pr\left[v'\middle|v_2\right]}\le\exp\left(\frac{\varepsilon_2\cdot\left|v_1-v_2\right|_1}{\Delta_{LP}}\right)\le\exp\left(\frac{\varepsilon_2\cdot\Delta_{LP}}{\Delta_{LP}}\right)=\exp\left(\varepsilon_2\right) \tag{S5}$$

Theorem 4.2 is proved.

Theorem 4.3. The mean squared error (MSE) of the Value Perturbation algorithm I (VP_LP) is $2\cdot\left(L/\varepsilon_2\right)^2$.

Proof:

The MSE of VP_LP can be defined as:

$$ErrorMSE_{VP\_LP}=\mathrm{E}\left[\left|v'-v\right|_2^2\right] \tag{S6}$$

Since the mean of the added noise is 0, the variance is:

$$D(x)=E\left(x^2\right)-E^2\left(x\right)=2\cdot b^2=2\cdot\left(\frac{\Delta_{LP}}{\varepsilon_2}\right)^2 \tag{S7}$$

From Eq (S7) we can get:

$$E\left(x^2\right)=2\cdot\left(\frac{\Delta_{LP}}{\varepsilon_2}\right)^2 \tag{S8}$$

As a result,

$$ErrorMSE_{VP\_LP}=\mathrm{E}\left[\left|v'-v\right|_2^2\right]=E\left[\left|Lap\left(\frac{\Delta_{LP}}{\varepsilon_2}\right)\right|^2\right]=2\cdot\left(\frac{\Delta_{LP}}{\varepsilon_2}\right)^2 \tag{S9}$$

Obviously, the MSE of VP_LP is directly proportional to the sensitivity $L_c=\left|IC_c\right|$, and we can conclude that:

$$ErrorMSE_{VP\_LP}=2\cdot\left(\frac{L_c}{\varepsilon_2}\right)^2 \tag{S10}$$

Theorem 4.3 is proved.

Theorem 4.4. VP_EM satisfies ε_2 - local differential privacy:

$$\frac{\Pr[VP\left(v_1\right)=v^*]}{\Pr[VP\left(v_2\right)=v^*]}\le e^{\varepsilon_2} \tag{S11}$$

Proof:

Let the two different inputs of VP be v_1, v_2, and the probability ratio of v^* returned by VP_EM is as follows:

$$\frac{\Pr[VP(v_1)=v^*]}{\Pr[VP(v_2)=v^*]} = \frac{\exp\left(\frac{\varepsilon_2 \cdot u_v(v_1,v^*)}{2}\right) \Big/ \sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_1,v^*)}{2}\right)}{\exp\left(\frac{\varepsilon_2 \cdot u_v(v_2,v^*)}{2}\right) \Big/ \sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_2,v^*)}{2}\right)}$$
$$= \underbrace{\frac{\exp\left(\frac{\varepsilon_2 \cdot u_v(v_1,v^*)}{2}\right)}{\exp\left(\frac{\varepsilon_2 \cdot u_v(v_2,v^*)}{2}\right)}}_{(1)} \cdot \underbrace{\frac{\sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_2,v^*)}{2}\right)}{\sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_1,v^*)}{2}\right)}}_{(2)} \tag{S12}$$

For part 1 of Eq (S12), it is observed that:

$$\frac{\exp\left(\varepsilon_2 \cdot u_v(v_1,v^*)\right)}{\exp\left(\varepsilon_2 \cdot u_v(v_2,v^*)\right)} = \exp\left(\frac{\varepsilon_2 \cdot \left(u_v(v_1,v^*) - u_v(v_2,v^*)\right)}{2}\right)$$
$$\le \exp\left(\frac{\varepsilon_2 \cdot \triangle u_v}{2}\right) \tag{S13}$$

Because $\triangle u_v \le 1$, part 1 from Eq (S12) is proved:

$$\frac{\exp\left(\varepsilon_2 \cdot u_v(v_1,v^*)\right)}{\exp\left(\varepsilon_2 \cdot u_v(v_2,v^*)\right)} \le \exp\left(\frac{\varepsilon_2}{2}\right) \tag{S14}$$

Next, for the part 2 of Eq (S12):

$$\frac{\sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_2,v^*)}{2}\right)}{\sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_1,v^*)}{2}\right)} = \frac{\sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_2,v^*) + \varepsilon_2 \cdot u_v(v_1,v^*) - \varepsilon_2 \cdot u_v(v_1,v^*)}{2}\right)}{\sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_1,v^*)}{2}\right)} \tag{S15}$$

$$\frac{\varepsilon_2 \cdot u_v(v_2,v^*)}{2} - \frac{\varepsilon_2 \cdot u_v(v_1,v^*)}{2} \le \frac{\varepsilon_2}{2} \cdot \left|u_v(v_2,v^*) - u_v(v_1,v^*)\right| \le \frac{\varepsilon_2}{2} \cdot \Delta u_v \le \frac{\varepsilon_2}{2} \tag{S16}$$

We apply Eq (S16) to Eq (S12) to get the conclusion of part 2 of Eq (S12):

$$\frac{\sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_2,v^*)}{2}\right)}{\sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_1,v^*)}{2}\right)} \le \exp\left(\frac{\varepsilon_2}{2}\right) \cdot \frac{\sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_1,v^*)}{2}\right)}{\sum_{y\in[0,l]} \exp\left(\frac{\varepsilon_2 \cdot u_v(v_1,v^*)}{2}\right)} = \exp\left(\frac{\varepsilon_2}{2}\right) \tag{S17}$$

By combining part 1 and part 2 we prove Theorem 4.4.

$$\frac{\Pr[VP(v_1)=v^*]}{\Pr[VP(v_2)=v^*]} \le \exp\left(\frac{\varepsilon_2}{2}\right)\cdot\exp\left(\frac{\varepsilon_2}{2}\right)=\exp(\varepsilon_2) \quad \text{(S18)}$$

Theorem 4.4 is proved.

Theorem 4.5. The MSE of VP_EM is:

$$ErrorMSE_{VP\_EM} = E\left(|v-y|^2\right)=\sum_{y=0}^{v} p_y\cdot\left(y^2-2\cdot v\cdot y\right)+v^2 \quad \text{(S19)}$$

and when $\forall y_1, y_2 \in [0,v], p_{y_1}=p_{y_2}$, i.e., when all sampling probabilities are the same, $ErrorMSE_{VP\_EM}$ reaches the maximum upper bound:

$$ErrorMSE_{VP\_EM} \le \frac{v\cdot(2\cdot v+1)}{6} \quad \text{(S20)}$$

Proof:

The real length of the set value data for user u is v, and the probability of the result y is:

$$p_y=\Pr\left[VP\_EM(v)=y\right]=\exp\left(\frac{\varepsilon_2\cdot u_v(y,v)}{2}\right)\Big/\Omega_v \quad \text{(S21)}$$

The exception of y is $E(y)$, and therefore the MSE of y can defined as follows:

$$ErrorMSE_{VP\_EM} = E\left(|v-y|^2\right)=E\left(y^2\right)-2\cdot v\cdot E(y)+v^2 \quad \text{(S22)}$$

y is a discrete random variable based on the definition for expectation of discrete random variables $E\left(y^2\right)$ and $E(y)$ are defined below:

$$E\left(y^2\right)=\sum_{y=0}^{v}\left(y^2\cdot p_y\right), E(y)=\sum_{y=0}^{v}\left(y\cdot p_y\right) \quad \text{(S23)}$$

Therefore,

$$\begin{aligned} ErrorMSE_{VP\_EM} &= \sum_{y=0}^{v}\left(y^2\cdot p_y\right)-2\cdot v\cdot\sum_{y=0}^{v}\left(y\cdot p_y\right)+v^2 \\ &= \underbrace{\sum_{y=0}^{v} p_y\cdot\left(y^2-2\cdot v\cdot y\right)}_{*}+v^2 \end{aligned} \quad \text{(S24)}$$

where $0<p_y<1$, $0\le y\le v$, set $g(y)=y^2-2\cdot v\cdot y$. Then, $-v^2\le g(y)\le 0$, and the following formula holds:

$$g(y_1)\ge g(y_2), v\ge y_2>y_1\ge 0 \quad \text{(S25)}$$

$g(y)$ monotonically declines under the range $0\le y\le v$. Part * from Eq (S24) is less than or equal to 0, the problem of getting the maximum value of function $ErrorMSE_{VP\_EM}$ can be transformed into the problem of getting the minimum value of part *, and the linear function of variable p_y can be defined as follows:

$$f(p_y)=\sum_{y=0}^{v} p_y\cdot g(y) \quad \text{(S26)}$$

when $\varepsilon_2 = 0$, all of p_y is equal, i.e., $\forall y_1, y_2 \in [0, v], p_{y_1} = p_{y_2} = \frac{1}{v+1}$, which means that the value of y is completely random with independent of utility function u_v, and this situation has the strongest privacy. When $\varepsilon_2 > 0$, $y_1 \le y_2$, $p_{y_1} \le p_{y_2}$, and on account of $f(p_y)$ being a monotonically increasing function, we can conclude that when $\forall y_1, y_2 \in [0, v], p_{y_1} = p_{y_2} = \frac{1}{v+1}$, $f(p_y)$ takes the minimum value, and $ErrorMSE_{VP\_EM}$ takes the maximum value as follows:

$$\begin{aligned} ErrorMSE_{VP\_EM} &\le \sum_{y=0}^{v} p_y \cdot \left(y^2 - 2 \cdot v \cdot y\right) + v^2 \\ &= \frac{1}{v+1} \cdot \left(\sum_{y=0}^{v} y^2 - 2 \cdot v \cdot \sum_{y=0}^{v} y\right) + v^2 \\ &= \frac{1}{v+1} \cdot \left(\frac{v \cdot (v+1) \cdot (2 \cdot v + 1)}{6} - 2 \cdot v \cdot \frac{v \cdot (v+1)}{2}\right) + v^2 \\ &= \frac{v \cdot (2 \cdot v + 1)}{6} \end{aligned} \tag{S27}$$

Theorem 4.5 is proved.

Theorem 4.6. RS satisfies ε_2 - local differential privacy.

Proof: The proof procedure is the same as Theorem 4.4.

Theorem 4.7.

Set $y = |t_c \cap t'_c|$ is the intersection size of t_c and t'_c, the probability is p_y, and the MSE of y is:

$$ErrorMSE_{RS} = E\left(|v - y|^2\right) = \sum_{y=0}^{v} p_y \cdot \left(y^2 - 2 \cdot v \cdot y\right) + v^2 \tag{S28}$$

where $v' = |t'_c|$, $v = |t_c|$, and $r = \min\{v, v'\}$. When $p_y = \frac{1}{1+r}$, $ErrorMSE_{RS}$ reaches the maximum upper bound:

$$ErrorMSE_{RS} \le \frac{v \cdot (v+1)}{1+r} \cdot \left(\frac{1 - 4 \cdot v}{6}\right) + v^2 \tag{S29}$$

In particular, where r=0, $ErrorMSE_{RS}$ reaches the maximum upper bound: $v \cdot (v+1) \cdot \left(\frac{1 - 4 \cdot v}{6}\right) + v^2$.

Proof:

v' is the value perturbation count, the real count is v, the length of the sub-domain under category c is L_c, the true set-valued data is t_c, the resulting itemset is s_y, and the length of the intersection (that is, the length of the retained data) is $y = |t_c \cap s_y|$. Then, the MSE of y is:

$$ErrorMSE_{RS} = E\left(|v - y|^2\right) = E\left(y^2\right) - 2 \cdot v \cdot E(y) + v^2 \tag{S30}$$

It can be seen that the possible values of v' are $[0, L_c]$, and set $r = \min\{v, v'\}$ is the maximum possible intersection count. Then, the value of y may be $[0, r]$. The probability of y is:

$$p_y = \underbrace{C_{v'}^{y} \cdot C_{L_c - y}^{v' - y}}_{*} \cdot \underbrace{\frac{\exp\left(\varepsilon_3 \cdot u'_{itemset}\left(s_y, t_c\right)/2\right)}{\Omega_{itemset}}}_{**} \tag{S31}$$

Part ** of Eq (S31) is the probability y. The number of possible s_y is $C_r^y \cdot C_{L_c-y}^{v'-y}$, and s_y means all the candidate itemset satisfies $y=|t_c \cap s_y|$ under category c, which is the * part of p_y. Intuitively, the * part of p_y is equivalent to dividing all candidate itemsets of length v' into r+1 subspaces. The range of intersection length y between the candidate set and original data in each subspace is [0, r], and p_y is the probability of $y=|t_c \cap s_y|$ in the subspace.

Similar to Theorem 4.6, when the probability of all subspaces is the same, i.e., $p_y = 1/(1+r)$, the upper bound of $ErrorMSE_{RS}$ is:

$$\begin{aligned} ErrorMSE_{RS} &= E\left(|v-y|^2\right) \\ &= \sum_{y=0}^{k} p_y \cdot \left(y^2 - 2\cdot v\cdot y\right) + v^2 \\ &\le \frac{1}{1+r}\cdot\sum_{y=0}^{r}\left(y^2-2\cdot v\cdot y\right)+v^2 \\ &= \frac{1}{1+r}\cdot\left(\frac{v\cdot(v+1)\cdot(2\cdot v+1)}{6} - 2\cdot v\cdot\frac{v\cdot(v+1)}{2}\right)+v^2 \\ &= \frac{v\cdot(v+1)}{1+r}\cdot\left(\frac{1-4\cdot v}{6}\right)+v^2 \end{aligned} \tag{S32}$$

When r=0, $ErrorMSE_{RS}$ reaches the upper bound: $v\cdot(v+1)\cdot\left(\frac{1-4\cdot v}{6}\right)+v^2$. Here, $r=\min\{v,v'\}$ is the maximum possible value of the intersection. This means that the result is the worst when there is no intersection between the returned itemset and the original set-value data.

Theorem 4-7 is proved.