SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Locally differentially private distributed algorithms for set intersection and union

Qiao XUE[1], Youwen ZHU[1,3*], Jian WANG[1], Xingxin LI[1] & Ji ZHANG[2]

[1]*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016*, China;*
[2]*Faculty of Health, Engineering and Sciences, The University of Southern Queensland, Toowoomba* 4350*, Australia;*
[3]*Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing* 210023*, China*

Appendix A The basic solution

Users first encode their multiset into a group of binary vectors. Then each user u_i perturbs all max binary vectors ($S_i = \{S_{i1}, \cdots, S_{imax}\}$) by randomized response mechanism. Since an item can repeat max times in multisets, the number of the same item in multisets of any two users can differ by at most max. Therefore, to make items in multisets satisfy ϵ-local differential privacy, each user should sanitize each binary vector with the privacy budget of $\frac{\epsilon}{max}$. Then, each user u_i's perturbed vectors $\widetilde{S}_i = \{\widetilde{S}_{i1}, \cdots, \widetilde{S}_{imax}\}$ are sent to the collector. Finally, the collector builds max perturbed matrixes $\widetilde{M}_1, ..., \widetilde{M}_{max}$. Each perturbed matrix $\widetilde{M}_t$ consists of the corresponding perturbed vectors of all users, i.e., $\widetilde{M}_t = [\widetilde{S}_{1t}^T, ..., \widetilde{S}_{nt}^T]$. From each perturbed matrix $\widetilde{M}_t$, the collector can derive $\hat{\boldsymbol{\rho}}_t = \{\hat{\rho}_t[1], ..., \hat{\rho}_t[l]\}$, the estimated percentage of '1' in every corresponding column in each noise-free matrix $M_t = [S_{1t}^T, ..., S_{nt}^T]$, and gain an intermediate intersection AI_t. The intermediate intersection AI_t indicates that items in it can repeat at least t times in the final intersection I. Thus, if one item appears in the t'-th intermediate intersection $AI_{t'}$ at last (that is, this item does not appear in subsequent intermediate intersections $AI_{t'+1}$, ..., AI_{max}), t' will be the multiplicity of the item in I. Hence, through finding the last intermediate intersection where each item occurs, the corresponding multiplicity of each item in I can be determined and thus I are achieved.

However, the naive approach suffers from two disadvantages obviously. For one thing, its communication complexity of each user is higher, $\mathcal{O}(max \times l)$. For another, the basic method divides privacy budget into max parts, and each portion of the privacy budget is just $\frac{\epsilon}{max}$, which leads to heavy perturbation and low accuracy.

Appendix B The proof of theorems

Theorem 1. For one item which is in the t'-th and the $(t'+1)$-th true intermediate intersections, the probability that the item is excluded from these two estimated intersection $AI_{t'}, AI_{t'+1}$ by mistake is less than $e^{\frac{-4\epsilon^\epsilon}{(\epsilon^\epsilon+1)^2}}$.

proof : Without loss of generality, suppose that the item is the j-th distinct item in the universal multiset. According to the Algorithm C1, the item can be excluded from the intersection $AI_{t'}$ if

$$\hat{\rho}_{t'}[j] < 1 - \sqrt{\frac{\frac{1}{4} - (\frac{1}{2} - p_{11})^2}{(p_{00} + p_{11} - 1)^2 n'}}. \tag{B1}$$

While $p = p_{00} = p_{11}$, equation (B1) can be simplified to the following equation,

$$\frac{p-1}{2p-1} + \frac{\widetilde{\rho}_{t'}[j]}{2p-1} < 1 - \frac{\sqrt{(1-p)p}}{(2p-1)\sqrt{n'}},$$

where n' is the number of the users who participate in the calculation of $AI_{t'}$.

Then, the probability of the item being excluded is,

$$\begin{aligned}
&Pr\{\frac{p-1}{2p-1} + \frac{\widetilde{\rho}_{t'}[j]}{2p-1} < 1 - \frac{\sqrt{(1-p)p}}{(2p-1)\sqrt{n'}}\} \\
=&Pr\{\sqrt{n'}(p-1) + \sqrt{n'}\widetilde{\rho}_{t'}[j] < (2p-1)\sqrt{n'} - \sqrt{(1-p)p}\} \\
=&Pr\{\widetilde{\rho}_{t'}[j] < p - \frac{\sqrt{(1-p)p}}{\sqrt{n'}}\},
\end{aligned}$$

* Corresponding author (email: zhuyw@nuaa.edu.cn)

Besides,

$$\widetilde{\rho}_{t'}[j] = \frac{1}{n'}\sum_{i=1}^{n'} \widetilde{S}_{it'}[j],$$

$$\mathbb{E}(\widetilde{\rho}_{t'}[j]) = \frac{1}{n'}\sum_{i=1}^{n'} \mathbb{E}(\widetilde{S}_{i}[j]) = p.$$

Thus,

$$\begin{aligned} &Pr\{\widetilde{\rho}_{t'}[j] < p - \frac{\sqrt{(1-p)p}}{\sqrt{n'}}\} \\ =&Pr\{\mathbb{E}(\widetilde{\rho}_{t'}[j]) - \widetilde{\rho}_{t'}[j] > \frac{\sqrt{(1-p)p}}{\sqrt{n'}}\}. \end{aligned}$$

Applying the Chernoff-Hoeffding bound, we have

$$\begin{aligned} &Pr\{\frac{p-1}{2p-1} + \frac{\widetilde{\rho}_{t}[j]}{2p-1} < 1 - \frac{\sqrt{(1-p)p}}{(2p-1)\sqrt{n'}}\} \\ =&Pr\{\mathbb{E}(\widetilde{\rho}_{t'}[j]) - \widetilde{\rho}_{t'}[j] > \frac{\sqrt{(1-p)p}}{\sqrt{n'}}\} \\ \leqslant&e^{-2(\frac{\sqrt{(1-p)p}}{\sqrt{n'}})^2 n'} \\ =&e^{\frac{-2e^{\epsilon}}{(e^{\epsilon}+1)^2}}. \end{aligned}$$

Determining whether the item is included in $AI_{t'}$ is independent of determining whether it is included in $AI_{t'+1}$, thus the probability of the item being excluded from both $AI_{t'}$ and $AI_{t'+1}$ by mistake is less than $e^{\frac{-4e^{\epsilon}}{(e^{\epsilon}+1)^2}}$. For instance, when $\epsilon = 0.5$, the probability is less than 0.15. The proof is completed.

Theorem 2. The perturbation method for multisets of users satisfies ϵ-local differential privacy.

proof : In the perturbation protocol, each user needs to pick a value T_i from $\{1, ..., max\}$ firstly, and then perturb his T_i-th vector. Given any two users $u_i, u_{i'}$ and the j-th distinct item in U, the probability of two users generating the same value $T \in \{1, ..., max\}$ (*i.e.*, $T_i = T_{i'} = T$) and the same value in j-th bit of the perturbed vector $\widetilde{Ms} \in \{0, 1\}$ (*i.e.*, $\widetilde{S}_{iT}[j] = \widetilde{S}_{i'T}[j] = \widetilde{Ms}$) is

$$\begin{aligned} &\frac{Pr(T, \widetilde{Ms}|S_{i1}[j], ..., S_{imax}[j])}{Pr(T, \widetilde{Ms}|S_{i'1}[j], ..., S_{i'max}[j])} \\ =&\frac{\frac{1}{max}Pr(\widetilde{Ms}|S_{iT}[j])}{\frac{1}{max}Pr(\widetilde{Ms}|S_{i'T}[j])} \\ =&\frac{Pr(\widetilde{Ms}|S_{iT}[j])}{Pr(\widetilde{Ms}|S_{i'T}[j])} \leqslant e^{\epsilon}. \end{aligned} \tag{B2}$$

Now, we use R' to denote the perturbation method in multiset schemes. We can get,

$$\frac{Pr(R(S_{i1}[j], ..., S_{imax}[j]) = (T, \widetilde{Ms}))}{Pr(R(S_{i'1}[j], ..., S_{i'max}[j]) = (T, \widetilde{Ms}))} \leqslant e^{\epsilon}.$$

and the perturbation method R is ϵ-local differential privacy.

The items are homogeneous, thus other distinct items can be proved in the same way. Therefore, the independent items in the perturbed vector of each user can satisfy ϵ-LDP, and then users' vectors (multisets) can be protected under local differential privacy, which completes the proof.

Appendix C Algorithm about Multiset Intersection Estimation

Algorithm C1 Multiset Intersection Estimation (MIE)

Require: uploaded data from all users $\widetilde{D}$=$\{(\widetilde{S}_{1T_1}, T_1), ..., (\widetilde{S}_{nT_n}, T_n)\}$, the privacy budget ϵ, the maximal multiplicity max.

Ensure: multiset intersection I.

for $t = 1$ to max **do**
  initialize $\widetilde{M}_t = [\,]$ /**** It is an empty matrix***/
  initialize $AI_t = \emptyset$, $\tilde{\boldsymbol{\rho}}_t = \mathbf{0} \in \{0\}^l$, $\hat{\boldsymbol{\rho}}_t = \mathbf{0} \in \{0\}^l$
end for
initialize $I = \emptyset$
for $i = 1$ to n **do**
  add u_i's perturbed vector $\widetilde{S}_i$ to the T_i-th matrix $\widetilde{M}_{T_i}$
end for
$p = \frac{e^\epsilon}{e^\epsilon+1}$
$error(n) = \sqrt{\frac{\frac{1}{4}-(-p+\frac{1}{2})^2}{(2p-1)^2 n}}$
for $t = 1$ to max **do**
  for $j = 1$ to l **do**
    calculate the percentage of '1' in the j-th column of the perturbed matrix $\widetilde{M}_t$, $\tilde{\boldsymbol{\rho}}_t[j]$
    estimate the percentage of '1' in the j-th column of corresponding noise-free matrix M_t, $\hat{\boldsymbol{\rho}}_t[j] = \frac{p-1}{2p-1} + \frac{\tilde{\boldsymbol{\rho}}_t[j]}{2p-1}$
    n'=size(M_t) /**** The number of rows of M_t***/
    if $\hat{\boldsymbol{\rho}}_t[j] \geqslant 1 - error(n')$ **then**
      insert the j-th itme of U in AI_t
    end if
  end for
end for
for $j = 1$ to l **do**
  for $t = 1$ to $max - 1$ **do**
    if $j \notin AI_t$ && $j \notin AI_{t+1}$ **then**
      the j-th item of U repeats $t - 1$ times in I
    end if
  end for
end for
return I
end function

Appendix D Figures

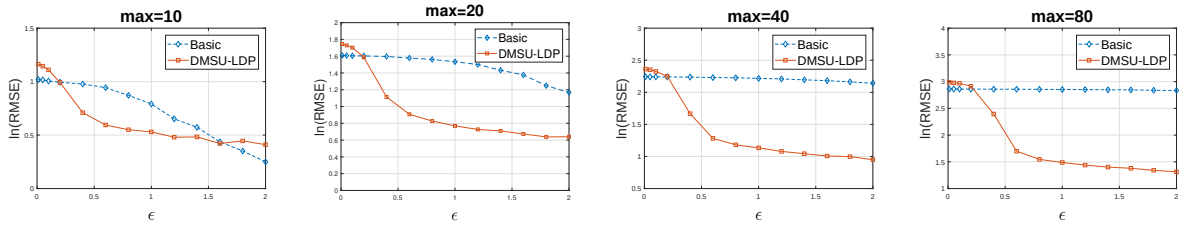

Figure D1 Error in estimated union of two different schemes on ***Accidents*** while the private budget ϵ ranges from 0.01 to 2 and the max varies from 10 to 80 .

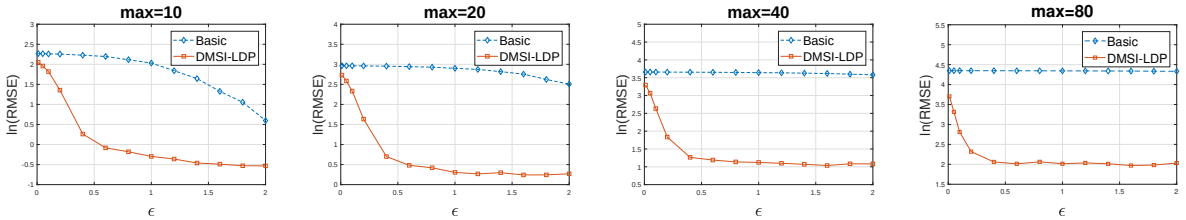

Figure D2 Error in estimated intersection of two different schemes on ***Accidents*** while the private budget ϵ ranges from 0.01 to 2 and the max varies from 10 to 80 .

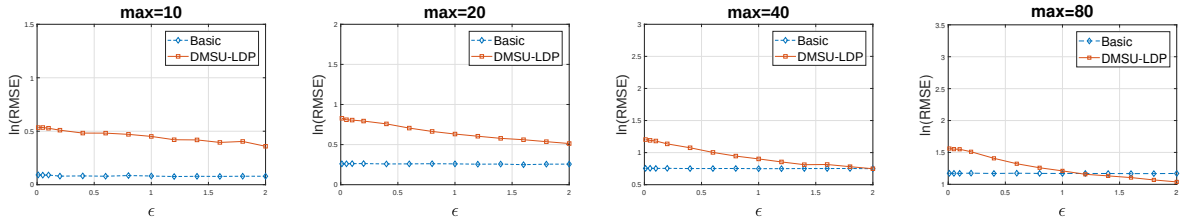

Figure D3 Error in estimated union of two different schemes on ***TK*** while the private budget ϵ ranges from 0.01 to 2 and the max varies from 10 to 80.

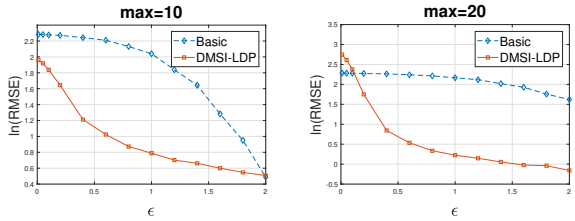

Figure D4 Error in estimated intersection of two different schemes on ***TK*** while the private budget ϵ ranges from 0.01 to 2 and the max ranges from 10 to 20.