Truth Inference on Sparse Crowdsourcing Data with Local Differential Privacy
IEEE BIG DATA '18

Haipei Sun\textsuperscript{1}        Boxiang Dong\textsuperscript{2}        Hui (Wendy) Wang\textsuperscript{1}
Ting Yu\textsuperscript{3}          Zhan Qin\textsuperscript{4}

\textsuperscript{1}Stevens Institute of Technology
Hoboken, NJ

\textsuperscript{2}Montclair State University
Montclair, NJ

\textsuperscript{3}Qatar Computing Research Institute
Doha, Qatar

\textsuperscript{4}The University of Texas at San Antonio
San Antonio, Texas
Data curator releases tasks on a crowdsourcing platform.
Crowdsourcing

- Data curator releases tasks on a crowdsourcing platform.
- The workers provide their answers to these tasks in exchange for a reward.
Collecting answers from individual workers may pose potential privacy risks.

- Crowdsourcing-related applications collect sensitive personal information from workers.
- By using a sequence of surveys, a data curator (DC) could potentially determine the identities of workers.
Differential privacy (DP) provides rigorous privacy guarantee.

However, classical DP requires a trusted data curator to publish privatized statistical information.
Local Differential Privacy

Local differential privacy (LDP) is the state-of-the-art approach for privacy-preserving data collection.

Before sending the answer to the data curator, each worker perturbs his/her private data locally.

\[ \hat{x}_1 = x_1 + \xi_1 \]
\[ \hat{x}_2 = x_2 + \xi_2 \]
\[ \hat{x}_m = x_m + \xi_m \]

\[ f(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m) \]
Challenges I - Data Sparsity

- Most workers only provide answers to a very small portion of the tasks.
- We use NULL to represent the answer if a worker does not provide response for a specific task.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Workers</th>
<th># of Tasks</th>
<th>Average Sparsity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web (^1)</td>
<td>34</td>
<td>177</td>
<td>0.705882</td>
</tr>
<tr>
<td>AdultContent (^2)</td>
<td>825</td>
<td>11,040</td>
<td>0.993666</td>
</tr>
</tbody>
</table>

- NULL values should also be protected.
- Careless perturbation of NULL values may significantly alter the original answer distribution.

---

\(^1\) [http://dbgroup.cs.tsinghua.edu.cn/ligl/crowddata/](http://dbgroup.cs.tsinghua.edu.cn/ligl/crowddata/)
Challenges II - Data Utility

• Truth inference estimates the true results from answers provided by workers of different quality.
• Most truth inference algorithms iterate until convergence.
• We aim to preserve the accuracy of truth inference on the perturbed worker answers, even a slight amount of initial noise in the worker answers may be propagated during iterations.
Our Contributions

Extension to Existing Approaches

- Laplace perturbation (LP) approach
- Randomized response (RR) approach
- Large expected error in the truth inference results

Novel Approach

We design a new matrix factorization (MF) perturbation algorithm to satisfy LDP, and guarantee small error.
Outline

1. Introduction
2. Related Work
3. Preliminaries
4. Perturbation Schemes
   - Laplace Perturbation (LP)
   - Randomized Response (RR)
   - Matrix Factorization (MF)
5. Experiments
6. Conclusion
Related Work

Local differential privacy

- Count, heavy hitters [HILM02, HIM02]
- Graph synthesis [QYY+17]
- Linear regression [NXY+16]

Privacy-preserving crowdsourcing

- Mutual information [KOV14]
- Truth discovery on complete data [LMS+18]

Differentially private recommendation

- Perturbation on categories [Can02, SJ14]
- Iterative factorization [SKSX18]
Preliminaries - Local Differential Privacy (LDP)

Definition ($\epsilon$-Local Differential Privacy)

A randomized privatization mechanism $\mathcal{M}$ satisfies $\epsilon$-local differential privacy ($\epsilon$-LDP) iff for any pair of answer vectors $\vec{a}$ and $\vec{a}'$ that differ at one cell, we have:

$$\forall \vec{z}_p \in \text{Range} (\mathcal{M}): \frac{Pr[\mathcal{M}(\vec{a}) = \vec{z}_p]}{Pr[\mathcal{M}(\vec{a}') = \vec{z}_p]} \leq e^\epsilon,$$

where $\text{Range}(\mathcal{M})$ denotes the set of all possible outputs of the algorithm $\mathcal{M}$. 
Preliminaries - Truth Inference

- Associated each worker with a quality.
- For each task, estimate the truth by taking the weighted average of the worker answers.
- For each worker, estimate the quality by measuring the difference between his answers and the estimated truth.

\[
\hat{\mu}_j = \frac{\sum_{w_i \in W_j} q_i \times a_{i,j}}{\sum_{w_i \in W_j} q_i}
\]

\[
\sigma_i = \sqrt{\frac{1}{|T_i|} \sum_{t_j \in T_i} (a_{i,j} - \hat{\mu}_j)^2}
\]
Iteratively updating the estimated truth and worker quality until convergence [LLG+14].

**Algorithm 1 Truth inference**

**Require:** The workers’ answers \( \{a_{i,j}\} \)

**Ensure:** The estimated true answer (i.e., the truth) of tasks \( \{\hat{\mu}_j\} \) and the quality of workers \( \{q_i\} \)

1: Initialize worker quality \( q_i = 1/m \) for each worker \( W_i \in \mathcal{W} \);
2: while the convergence condition is not met do
3:  Estimate \( \{\hat{\mu}_j\} \);
4:  Estimate \( \{q_i\} \);
5: end while
6: return \( \{\hat{\mu}_j\} \) and \( \{q_i\} \);
Given $M \in \mathbb{R}^{m \times n}$, find $U \in \mathbb{R}^{m \times d}$ and $V \in \mathbb{R}^{n \times d}$ s.t. 
$L(M, U, V) = \sum_{(i,j) \in \Omega} (M_{i,j} - \vec{u}_i^T \vec{v}_j)^2$ is minimized.

$M_{i,j}$, can be approximated by the inner product of $\vec{u}_i$ and $\vec{v}_j$, i.e., $\vec{u}_i^T \vec{v}_j$. 
**Problem Statement**

**Input** A set of answers \( \{ W_i \} \) and their answer vectors \( A = \{ \vec{a}_i \} \), and a privacy parameter \( \epsilon \)

**Output** The perturbed answer vectors \( A^P = \{ \mathcal{M}(\vec{a}_i) | \forall \vec{a}_i \in A \} \)

**Requirement**

- **Privacy:** \( A^P \) satisfies \( \epsilon \)-LDP.
- **Utility:** Accurate truth inference results from \( A^P \), i.e., minimize

\[
MAE(A^P) = \frac{\sum_{T_j \in T} |\mu_j - \hat{\mu}_j|}{n}.
\]
Laplace Perturbation (LP)

**Step 1** Replace NULL values with some value in the answer domain $\Gamma$.

$$g(a_{i,j}) = \begin{cases} v & a_{i,j} = NULL \\ a_{i,j} & a_{i,j} \neq NULL, \end{cases}$$

**Step 2** Add Laplace noise to each answer.

$$\mathcal{L}(\tilde{a}_i) = (g(a_{i,1}) + \text{Lap}(\frac{|\Gamma|}{\epsilon}), g(a_{i,2}) + \text{Lap}(\frac{|\Gamma|}{\epsilon}), \ldots, g(a_{i,n}) + \text{Lap}(\frac{|\Gamma|}{\epsilon}))$$
Laplace Perturbation (LP)

**Theorem 1 (Expected MAE of LP)**

Given a set of answer vectors $A = \{\vec{a}_i\}$, let $A^P = \{\hat{a}_i\}$ be the answer vectors after applying LP on $A$. Then the expected error $E\left[\text{MAE}(A^P)\right]$ of the estimated truth on $A^P$ must satisfy that

$$
E\left[\text{MAE}(A^P)\right] \leq \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{m} (q_i \times e_{i,j}^{LP}),
$$

where $e_{i,j}^{LP} = (1 - s_i) \left( \phi_j + \frac{|\Gamma|}{\epsilon} \right) + s_i \left( \sigma_i \sqrt{\frac{2}{\pi}} + \frac{|\Gamma|}{\epsilon} \right)$, $\mu_j$ is the ground truth of task $T_j$, $\sigma_i$ is the standard error deviation of worker $W_i$, $s_i$ is the fraction of the tasks that $W_i$ returns non-NULL values, and $\phi_j$ is the deviation between $\mu_j$ and the expected value $E(v)$ of $v$. 
Laplace Perturbation (LP)

Simple Setting

• \( q_i = \frac{1}{m}, \sigma_i = 1 \), i.e., all workers have the same quality.
• \( \mu_j = 1 \), i.e., all ground truths are 1.
• \( s_i = 0.1 \), i.e., 10% answers are not NULL.
• \( |\Gamma| = 10 \).
• \( \epsilon = 1 \).

Expected Error

\[
E \left[ MAE(A^P) \right] \leq 14.13
\]
Randomized Response (RR)

• Add NULL to the answer domain \( \Gamma \).
• For each answer \( a_{i,j} \), apply randomized response.

\[
\forall y \in \Gamma, \ Pr[\mathcal{M}(a_{i,j}) = y] = \begin{cases} 
\frac{e^\epsilon}{|\Gamma|+e^\epsilon} & \text{if } y = a_{i,j} \\
\frac{1}{|\Gamma|+e^\epsilon} & \text{if } y \neq a_{i,j}
\end{cases}
\]

Each original answer either
• remains unchanged in with probability \( \frac{e^\epsilon}{|\Gamma|+e^\epsilon} \), or
• is replaced with a different value with probability \( \frac{1}{|\Gamma|+e^\epsilon} \).
Randomized Response (RR)

Theorem 2 (Expected MAE of RR)

Given a set of answer vectors \( A = \{ \vec{a}_i \} \), let \( A^P = \{ \hat{a}_i \} \) be the answer vectors after applying RR on \( A \). Then the expected error \( E \left[ \text{MAE}(A^P) \right] \) of the estimated truth on \( A^P \) must satisfy that

\[
E \left[ \text{MAE}(A^P) \right] \leq \frac{1}{n} \sum_{j=1}^{n} \frac{\sum_{W_i \in W_j} q_i \times e_{i,j}^{RR}}{\sum_{W_i \in W_j} q_i},
\]

where

\[
e_{i,j}^{RR} = (1 - s_i) \left| \mu_j - \sum_{y \in \Gamma} y \frac{1}{\epsilon + |\Gamma|} \right| + \sum_{x \in \Gamma} s_i \mathcal{N}(x; \mu_j, \sigma_i) \left| \mu_j - \sum_{y \in \Gamma} y P_{xy} \right|,
\]

\( s_i \) is the fraction of tasks that worker \( W_i \) returns non-NULL values, and \( P_{xy} \) is the probability that value \( x \) is replaced with \( y \).
Randomized Response (RR)

Simple Setting

- \( q_i = \frac{1}{m} \), \( \sigma_i = 1 \), i.e., all workers have the same quality.
- \( \mu_j = 0 \), i.e., all ground truths are 1.
- \( s_i = 0.1 \), i.e., 10% answers are not NULL.
- \( \Gamma = [0, 9] \).
- \( \epsilon = 1 \).

Expected Error

\[
E \left[ \text{MAE}(A^P) \right] \leq 3.551
\]
Matrix Factorization (MF)

- DC randomly generates the task profile matrix $V \in \mathbb{R}^{n \times d}$, and sends both $V$ and the tasks $\mathcal{T}$ to the workers.
Matrix Factorization (MF)

- DC randomly generates the task profile matrix $V \in \mathbb{R}^{n \times d}$, and sends both $V$ and the tasks $T$ to the workers.
- Every worker gets the answers $\vec{a}_i$, and returns the differentially private answer profile vector $\vec{u}_i$. 

\[
\{ \vec{a}_i = \vec{u}_i V \}
\]
Instead of directly adding noise to $\vec{u}_i$, we design a novel approach based on objective perturbation to reduce the distortion.

$$\vec{u}_i = \arg \min_{\vec{u}_i} L_{DP}(\vec{a}_i, \vec{u}_i, V).$$

$$L_{DP}(\vec{a}_i, \vec{u}_i, V) = \sum_{T_j \in T_i} (a_{i,j} - \vec{u}_i^T \vec{v}_j)^2 + 2\vec{u}_i^T \vec{\eta}_i,$$

where $\vec{\eta}_i = \{Lap(\frac{\Gamma}{\epsilon}), \ldots, Lap(\frac{\Gamma}{\epsilon})\}$ is a $d$-dimensional vector.
Matrix Factorization (MF)

Theorem 3 (LDP of MF)
The $MF$ mechanism guarantees $\epsilon$-LDP.
Matrix Factorization (MF)

Theorem 4 (Expected MAE of MF)

Given a set of answer vectors \( A = \{ \vec{a}_i \} \), let \( A^P = \{ \hat{a}_i \} \) be the answer vectors after applying MF on \( A \). The expected error \( E[\text{MAE}(A^P)] \) of estimated truth based on the answer vectors perturbed by the MF mechanism satisfies that:

\[
E[\text{MAE}(A^P)] \leq \tilde{q} m \left( \sqrt{\frac{2}{\pi}} + \frac{d|\Gamma|}{n\epsilon} \right),
\]

where \( \tilde{q} = \max_i \{ q_i \} \) and \( d \) is the factorization parameter.

**Property** The error bound is insensitive to answer sparsity.
Matrix Factorization (MF)

**Simple Setting**
- \( q_i = \frac{1}{m}, \sigma_i = 1 \), i.e., all workers have the same quality.
- \( \Gamma = [0, 9] \).
- \( \epsilon = 1 \).
- \( n = 1,000 \), i.e., 1,000 tasks.
- \( d = 100 \).

**Expected Error**

\[
E \left[ \text{MAE}(A^P) \right] \leq 1.8
\]
Experiments

Real-word Datasets

- Web dataset
  - 34 workers
  - 177 tasks
  - 0.7059 sparsity
- AdultContent dataset
  - 825 workers
  - 11,040 tasks
  - 0.9937 sparsity

Synthetic Dataset

Baseline 2-Layer approach [LMS+18]
Experiments

Error v.s. Privacy Budget

(a) sparsity = 0.9
(b) sparsity = 0.5

Synthetic dataset (2,000 workers, 200 tasks)

- MF always provides the smallest MAE.
- The accuracy provided by MF is not sensitive to the privacy budget.
Experiments

Error v.s. Answer Sparsity

(a) $\epsilon = 0.1$
(b) $\epsilon = 1.0$

Synthetic dataset (2,000 workers, 200 tasks)

- MF always provides the smallest MAE.
- The accuracy provided by MF is not sensitive to the data sparsity.
Experiments

Error v.s. Privacy Budget

(a) Web dataset (b) AdultContent dataset

Real-world datasets

- MF provides the lowest MAE for most cases.
We aim at protecting worker privacy with LDP guarantee while providing highly accurate truth inference results.

- Propose LP and RR to address sparsity in worker answers.
- Design MF that adds perturbation on objective functions.
- MF provides better data utility.

In the future, we aim at protecting task privacy.
References


[LLG+14] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, 8(4):425–436, 2014.


Thank you!

Questions?