Privacy in spatial crowdsourcing

A Framework for Protecting Worker Location Privacy in Spatial Crowdsourcing

CSCI 587 Nov 12 2014

Cyrus Shahabi
Motivation

Ubiquity of mobile users

6.5 billion mobile subscriptions, 93.5% of the world population [1]

Technology advances on mobiles

Smartphone's sensors. e.g., video cameras

Network bandwidth improvements

From 2.5G (up to 384Kbps) to 3G (up to 14.7Mbps) and recently 4G (up to 100 Mbps)

Spatial Crowdsourcing

- Crowdsourcing
  - Outsourcing a set of tasks to a set of workers

- Spatial Crowdsourcing
  - Crowdsourcing a set of spatial tasks to a set of workers.
  - Spatial task is related to a location e.g., taking pictures

Location privacy is one of the major impediments that may hinder workers from participation in SC
Problem Statement

Current solutions require the workers to disclose their locations to untrustworthy entities, i.e., SC-server.

A framework for protecting privacy of worker locations, whereby the SC-server only has access to data sanitized according to differential privacy.
Outline

❖ Background
❖ Privacy Framework
❖ Worker PSD (Private Spatial Decomposition)
❖ Task Assignment
❖ Experiments
Utility-Privacy Trade-off

Utility

100%

Privacy

0%
Related Work

- Pseudonymity (using fake identity)
  - e.g. fake identity + location == resident of the home

- K-anonymity model (not distinguish among other k records)
  identities are known
  the location k-anonymity fails to prevent the location of a subject
  being not identifiable
    all k users reside in the exact same location
  k-anonymity, do not provide rigorous privacy

- Cryptography
  such technique is computational expensive

=> not suitable for SC applications
Differential Privacy (DP)

DP ensures an adversary do not know from the sanitized data whether an individual is present or not in the original data

**ε-distinguishability** [Dwork’06]

A database produces transcript \( U \) on a set of queries. Transcript \( U \) satisfies \( \mathcal{E} \)-distinguishability if for every pair of sibling datasets \( D_1 \) and \( D_2 \), \(|D_1| = |D_2|\) and they differ in only one record, it holds that

\[
\ln \frac{\Pr[QS^{D_1} = U]}{\Pr[QS^{D_2} = U]} \leq \varepsilon
\]

DP allows only aggregate queries, e.g., count, sum.

**L₁-sensitivity:**

Given neighboring datasets \( D_1 \) and \( D_2 \), the sensitivity of query set \( QS \) is the maximum change in their query results

\[
\sigma(QS) = \max_{D_1, D_2} \sum_{i=1}^{q} |QS(D_1) - QS(D_2)|
\]

[Dwork’06] shows that it is sufficient to achieve \( \mathcal{E}\text{-DP} \) by adding random Laplace noise with mean

\[
\lambda = \sigma(QS) / \varepsilon
\]
Outline

- Background
- Privacy Framework
- Worker Private Spatial Decomposition
- Task Assignment
- Experiments
Privacy Framework

0. Workers send their locations to a trusted CSP

1. CSP releases a PSD according to $\varepsilon$. PSD is accessed by SC-server

2. SC-server receives tasks from requesters

3. When SC-server receives task $t$, it queries the PSD to determine a $GR$ that enclose sufficient workers. Then, SC-server initializes geocast communication to disseminate $t$ to all workers within $GR$

4. Workers confirm their availability to perform the assigned task

Workers trust SCP

Workers do not trust SC-server and requesters

Focus on private task assignment rather than post assignment
Design Goal and Performance Metrics

Protecting worker location may reduce the effectiveness and efficiency of worker-task matching, captured by following metrics:

**Assignment Success Rate (ASR):** measures the ratio of tasks accepted by workers to the total number of task requests

**Worker Travel Distance (WTD):** the average travel distance of all workers

**System Overhead:** the average number of notified workers (ANW). ANW affects both communication overhead required to geocast task requests and the computation overhead of matching algorithm
Outline

- Background
- Privacy Framework
- Worker PSD (Private Spatial Decomposition)
- Task Assignment
- Experiments
Adaptive Grid (Worker PSD)

Creates a coarse-grained, fixed size $m_1 \times m_1$ grid over data domain. Then issues $m_1^2$ count queries for each level-1 cell using $\varepsilon_1$

$$m_1 = \max \left( 10, \left\lceil \frac{1}{4} \sqrt{\frac{N \times \varepsilon}{k_2}} \right\rceil \right)$$

Partitions each level-1 cell into $m_2 \times m_2$ level-2 cells, $m_2$ is adaptively chosen based on noisy count $N'$ of level-1 cell

$$m_2 = \left\lceil \frac{1}{4} \sqrt{\frac{N' \times \varepsilon_2}{k_2}} \right\rceil$$

$$\varepsilon = \varepsilon_1 + \varepsilon_2$$
Customized AG

Expected #workers (noisy count) in level-2 cells

\[ \bar{n} = \frac{N'}{m_2^2} = \frac{k_2}{\epsilon_2} \]

Large \( n \) leads to high communication cost

\[ N' = 100 \]

<table>
<thead>
<tr>
<th>( \epsilon )</th>
<th>( \epsilon_2 )</th>
<th>( m_2 )</th>
<th>( \bar{n} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>0.1</td>
<td>0.05</td>
<td>1</td>
<td>100</td>
</tr>
</tbody>
</table>

😍 Original AG  \((k_2 = 5)\)

😊 Customized AG  \((k_2 = \sqrt{2}, p_h = 88\%)\)

Increase \( m_2 \) to decrease overhead, but only to the point where there is at least one worker in a cell

The probability that the real count is larger than zero:

\[ p_h = 1 - \frac{1}{2} \exp\left(-\frac{\text{count}_{\text{PSD}}}{1/\epsilon_2}\right) \]
Customized AG

- Original AG and Customized AG adapts to data distributions.
- Original AG minimizes overall estimation error of region queries while customized AG increases the number of 2nd level cells.

Yelp Dataset

Original AG

Customized AG
Outline

- Background
- Privacy Framework
- Worker PSD (Private Spatial Decomposition)
- Task Assignment
- Experiments
Analytical Utility Model

We define *Acceptance Rate* as a decreasing function of task-worker distance (e.g. linear, Zipian)

\[ p^a = F(d); \quad 0 \leq p^a \leq 1 \]

SC-server establishes an *Expected Utility* (*EU*) threshold, which is the targeted success rate for a task. *EU* > *p*^a^.

\[ X \] is a random variable for an event that a worker accepts a received task

\[ P(X = True) = p^a; \quad P(X = False) = 1 - p^a \]

Assuming \( w \) independent workers. *U* is the probability that at least one worker accepts the task

\[ X \sim Binomial\,(w, \, p_a) \]

\[ \Rightarrow U = 1 - (1 - p^a)^w \]
Acceptance Rate Functions

Acceptance Rate vs. distance and MTD

Utility (Monotonicity) vs. The number of workers for AR = 0.5
Geocast Region Construction

Determines a small region that contains sufficient workers

Greedy Algorithm (GDY)

1. Init $GR = \{\}$, max-heap $Q$ of candidates
   
   $Q = \{ \text{the cell that contains } t \}$

2. $c_i \leftarrow Q$

3. $U \leftarrow 1 - (1 - U)(1 - U_{c_i})$

4. If $U \geq EU$, return $GR$

5. $neighbors = \{c_i' \text{'s neighbors} \} - GR \cap MTD$

6. $Q = Q \cup neighbors$, Go to 2.
The number of workers can still be large with AG, especially when $\varepsilon_2$ small.

Allow *partial cell inclusion* on the lastly added cell $C_i$.
The more compact the GR, the lower the cost

**Digital Compactness Measurement [Kim’84]**

\[
DCM = \frac{\text{area}(GR)}{\text{area}(MIN \ BALL)}
\]

Measurement:

**Hop count** = \(\frac{\text{Farthest distance between two workers}}{2 \times \text{Communication range}}\)
Geocast Regions

A

B

C

D
Outline

• Background
• Privacy Framework
• Worker PSD (Private Spatial Decomposition)
• Task Assignment
• Experiments
Experimental Setup

• **Datasets**

<table>
<thead>
<tr>
<th>Name</th>
<th>#Tasks</th>
<th>#Workers</th>
<th>MTD (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gowalla</td>
<td>151,075</td>
<td>6,160</td>
<td>3.6</td>
</tr>
<tr>
<td>Yelp</td>
<td>15,583</td>
<td>70,817</td>
<td>13.5</td>
</tr>
</tbody>
</table>

• **Assumptions**
  – Gowalla and Yelp users are workers
  – Check-in points (i.e., of restaurants) are task locations

• **Parameter settings**
  \( \varepsilon = \{0.1, 0.4, 0.7, 1\} \)
  \( EU = \{0.3, 0.5, 0.7, 0.9\} \)
  \( MaxAR = \{0.1, 0.4, 0.7, 1\} \)

• 1000 random tasks x 10 seeds
GR Construction Heuristics (Gow.-Linear)

GDY = geocast (GREedy algorithm) + original Adaptive grid (AG)  [Qardaji’13]
G-GR = geocast + AG with customized GRanularity
G-PA = geocast  with PArtial cell selection + original Adaptive grid (AG)
G-GP = geocast  with Partial cell selection + AG with customized Granularity
Effect of Grid Size to ASR

Average ASR over all values of budget by varying $k_2$
Compactness-based Heuristics
(Yelp-Zipf)

HOP

ANW

G-GP-Pure
G-GP-Hybrid
G-GP-Compact

Eps=0.1  Eps=0.4  Eps=0.7  Eps=1

Eps=0.1  Eps=0.4  Eps=0.7  Eps=1
Overhead of Archiving Privacy (Gow.-Zipf)

ANW

WTD-FC

ASR

Privacy

Non-Privacy

Eps=0.1  Eps=0.4  Eps=0.7  Eps=1
Effect of Varying MAR
(Yelp-Linear)

ANW

WTD-FC

CELL

AR=0.1  AR=0.4  AR=0.7  AR=1
AR=0.1  AR=0.4  AR=0.7  AR=1
AR=0.1  AR=0.4  AR=0.7  AR=1

Eps=0.1  Eps=0.4  Eps=0.7  Eps=1
Eps=0.1  Eps=0.4  Eps=0.7  Eps=1
Eps=0.1  Eps=0.4  Eps=0.7  Eps=1
Effect of Varying EU (Yelp-Linear)
Demo

http://geocast.azurewebsites.net/geocast/

Usage Instruction
1. Cell service provider publishes dataset with privacy protection
2. SC-server selects a dataset to query
3. SC-server chooses algorithm parameter settings
4. Administrator performs geocast queries on map-based interface

https://www.youtube.com/watch?v=4zkiJ9gk79s
Conclusion

Introduced a novel privacy-aware framework in SC, which enables workers participation without compromising their location privacy.

Identified geocasting as a needed step to preserve privacy prior to workers consenting to a task.

Provided heuristics and optimizations for determining effective geocast regions that achieve high assignment success rate with low overhead.

Experimental results on real datasets shows that the proposed techniques are effective and the cost of privacy is practical.
References

Hien To, Gabriel Ghinita, Cyrus Shahabi. *A Framework for Protecting Worker Location Privacy in Spatial Crowdsourcing*. In Proceedings of the 40th International Conference on Very Large Data Bases (VLDB 2014)