Privacy Preservation for Location-Based Services Based on Attribute Visibility

Masanori Mano, Xi Guo, Tingting Dong, Yoshiharu Ishikawa
Outline

- Background
- Motivation
- Related work
- Overview of the approach
- Anonymization algorithm
- Experimental evaluation
- Conclusions and future work
Background
Location-Based Services (LBSs)

- LBSs are useful and popular
  - Provide services to mobile users according to their geographical locations
    - Show nearby cafés, gas-stations, restaurants, ...
    - Compute the best route to the destination
    - Send coupons provided by nearby restaurants
Technologies Supporting LBSs

- **Positioning technology:** obtain users’ locations
  - Example: GPS chips/satellites, cellphone triangulation, …

- **Networking technology:** access to Internet everywhere
  - Example: 3G, WiFi, …

- **Database technology:** develop colorful applications
Privacy Issue

- However, the LBS providers might be un-trusted or even adversaries
  - Identity (E.g., name, phone number, IP address, …)
  - Sensitive location (E.g., home, night club, clinic, …)
  - Malicious usage (E.g., keep and sell users’ logs, track users’ movements, …)
Protect Privacy

- Anonymizer, a trusted third party server
  - Place in-between users and LBS providers
  - Protect privacy by anonymizing users
  - Spatial cloaking [MobiSys03, VLDB06, WWW08]
Spatial Cloaking

- Anonymizer groups $k$ near users and send the group information to LBS providers
  - Prevent the adversary from identifying an individual with probability above $1/k$
  - Guarantee service quality by limiting the size of cloaked regions
Motivation
Personalized LBSs

- LBSs typically utilize user locations
  - Applications
    - Show restaurants nearby
    - Compute the best route to the destination
  - Protect privacy
    - Spatial cloaking

- Personalized LBSs utilize both locations and profiles
  - **Profile**: age, sex, occupation, …
  - Applications
    - Mobile shopping
    - Mobile advertising
  - Protect privacy?
Personalized LBS Example

- Location-based advertising (LBA)
  - Provide local advertisements to appropriate persons
    - Use location information to attract nearby users
    - Use profiles to avoid spam that make users unhappy

![Diagram showing personalized LBS example]

Jack, 60, M
Mary, 25, F
Bob, 8, M
Toy Coupons
Wine Coupons
Cosmetic Coupons
LBA Provider
However, the adversary can distinguish users

- Associate users with profiles by **watching** the target area
Our Idea to Protect Privacy

- Group the near users with similar profiles
- Reduce the identification probability
- Guarantee the quality of service (unchanged size of the cloaked region)
Related Work
Protect Privacy in LBSs

- In traditional LBSs [MobiSys03], [VLDB06], [WWW08], [TMC08]
  - Spatial cloaking
  - Construct cloaked regions that contain near users

- In personalized LBSs [MDM08]
  - Most anonymization methods do not consider users’ profiles
  - One exception is [MDM08], but it does not consider the attribute observability
    - Adversaries can associate profiles with users by watching
Personalized Anonymization

- Users specify their preferences of the attribute disclosure levels [SIGMOD06]
  - Static databases
  - Construct a hierarchical taxonomy for each attribute

Our work

- Spatial databases
  - Service request stream
  - Moving users
  - Hierarchical taxonomy
Details of the Approach
Attribute Observability

- **Observability** measures the easiness that adversaries can guess attribute values by observing
  - High observability
    - “Age”, “Sex”, …
  - Low observability
    - “Birthplace”, “Occupation” …
Personalized Anonymization

- Users specify their anonymization preferences
  - Attribute disclosure level (Lower level, disclose less)
  - Identification probability threshold
- According to the preferences, anonymizer construct cloaked regions and the anonymized profiles
Attribute Disclosure Level

- Generalize attribute values by hierarchical taxonomy

### Taxonomy of “Age”

<table>
<thead>
<tr>
<th>Level</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>any</td>
</tr>
<tr>
<td>Level 1</td>
<td>-19, 20-39, 40-</td>
</tr>
<tr>
<td>Level 2</td>
<td>20-29</td>
</tr>
<tr>
<td>Level 3</td>
<td>20-24, 25-29, 30-34, 35-39</td>
</tr>
</tbody>
</table>

### Anonymized Profiles

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Disclosure Level</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>23, Level 2</td>
<td>[20-29]</td>
<td>0.4</td>
</tr>
<tr>
<td>Mary</td>
<td>26, Level 1</td>
<td>[20-39]</td>
<td>0.5</td>
</tr>
<tr>
<td>Ann</td>
<td>22, Level 3</td>
<td>[20-24]</td>
<td>0.6</td>
</tr>
</tbody>
</table>
Identification Probability Threshold

- Identification probability \((Pr.)\)
  - The probability that the individual is identified
- Threshold \((T)\)
  - The highest probability permitted by the user

### Anonymized Profiles

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>(T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>[20-29]</td>
<td>0.4</td>
</tr>
<tr>
<td>Mary</td>
<td>[20-39]</td>
<td>0.5</td>
</tr>
<tr>
<td>Ann</td>
<td>[20-24]</td>
<td>0.6</td>
</tr>
</tbody>
</table>

### The Truth

- Pr. < 0.4: satisfied
- Threshold
  - Mary
  - Ann
  - Alice
Matching Degree

- The probability that a user can be related to an attribute value by watching
  - The probability is an empirical value
  - Describe the observability of an attribute value

![Diagram showing levels and matching degree](image)
Matching Degree Table

- Record all the matching degrees between users and nodes in the taxonomy tree
- Anonymizer owns the matching degree table

<table>
<thead>
<tr>
<th>ID</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[20-39]</td>
<td>[20-29]</td>
<td>[30-39]</td>
</tr>
<tr>
<td></td>
<td>[20-24]</td>
<td>[25-29]</td>
<td>[30-34]</td>
</tr>
<tr>
<td>0.88</td>
<td>0.88</td>
<td>0.00</td>
<td>0.54</td>
</tr>
<tr>
<td>1.00</td>
<td>0.90</td>
<td>0.10</td>
<td>0.38</td>
</tr>
<tr>
<td>0.79</td>
<td>0.79</td>
<td>0.00</td>
<td>0.56</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Calculate Identification Probability (cont.)

- Calculate the identification probabilities by looking up the matching degree table

<table>
<thead>
<tr>
<th>Age</th>
<th>Name</th>
<th>Pr1</th>
<th>Pr2</th>
</tr>
</thead>
<tbody>
<tr>
<td>[20-24]</td>
<td>Alice</td>
<td>0.54</td>
<td>0.34</td>
</tr>
<tr>
<td>[25-29]</td>
<td>Mary</td>
<td>0.52</td>
<td>0.38</td>
</tr>
</tbody>
</table>

\[
Pr1 = 0.54 \times 0.52 = 0.28 \\
Pr2 = 0.34 \times 0.38 = 0.13
\]

Matching Degree Table

<table>
<thead>
<tr>
<th>(u_i)</th>
<th>(Level 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[20-24]</td>
</tr>
<tr>
<td></td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>0.38</td>
</tr>
</tbody>
</table>

Identification Probability

\[
\text{Identification Probability} = \frac{Pr1}{Pr1 + Pr2} = \frac{0.28}{0.28 + 0.13} = 0.69
\]
Anonymization Algorithm
Anonymization Process

- **Input (sporadic user requests)**
  - Profile (name, age, ...)
  - Location (geographical coordinate)
  - Anonymization preference (disclosure level, threshold)

- **Construct candidate group**
  - The identification probability ($Pr.$) of each user should be lower than the threshold ($T$) permitted by her.
  - The cloaked region should be smaller than the maximum size specified by the service provider.

- **Output**
  - Candidate group

$$Probs. < Ts$$
$$Size < Limit$$
Temporal Information of User Requests

- **Starting time**
  - When the user requests the service

- **Duration**
  - How long the user is willing to wait

- **Deadline**
  - Starting time + Duration

[Diagram showing starting times and durations for users u1, u2, u3, and ellipses...]
Naïve Approach

- Process requests in the order of their deadlines
- When a candidate group is constructed successfully, output it immediately

Users ordered by deadlines: $u_1, u_2, u_3, u_4…$
Optimization Idea

- Wait for the appearance of a better candidate group until the earliest deadline came
- Six different approaches

Non-candidate

Candidate

Naïve

Deadline came

Optimization
Optimization Approaches (2/6)

- Deadline-based (candidate first)
  - Add the new user into the existing candidate groups
  - If no candidate group can merge it, construct new groups

- Lazy (non-candidate first)
  - Add the new user into the existing non-candidate groups to make the groups satisfying the thresholds
Optimization Approaches (4/6)

- **Many-first**: Output the candidate group containing the largest number of users.
- **Next-deadline-based**: Output the candidate group containing the next-earliest deadline user.
- **Avg-deadline-based**: Output the candidate group with the earliest average deadline.
- **Threshold-based**: Output the candidate group containing the lowest-threshold user.

![Diagram showing optimization approaches](image)
Experiments
## Settings

<table>
<thead>
<tr>
<th>Experimental parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>1000</td>
</tr>
<tr>
<td>Request frequencies</td>
<td>10 times/s (default)</td>
</tr>
<tr>
<td>Expiration duration (deadline)</td>
<td>10s ±10% (default)</td>
</tr>
<tr>
<td>Used attribute</td>
<td>Age</td>
</tr>
<tr>
<td>Age range</td>
<td>[20, 39]</td>
</tr>
<tr>
<td>Disclosure level</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>Threshold probability</td>
<td>0.3, 0.4, 0.5 (default)</td>
</tr>
<tr>
<td>Cloaked area size limit</td>
<td>1000 × 1000 (default)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evaluation criteria</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput</td>
<td>The number of users</td>
</tr>
<tr>
<td></td>
<td>successfully anonymized</td>
</tr>
<tr>
<td>Quality</td>
<td>The average disclosure</td>
</tr>
<tr>
<td></td>
<td>level</td>
</tr>
</tbody>
</table>
Varying Request Frequencies

Good throughput with the increase of frequencies

Bad throughput with the increase of frequencies

Throughput

Legend:
- 5 persons/sec
- 10 persons/sec
- 50 persons/sec
- 100 persons/sec
Varying Maximum Size of Cloaked Region

Good throughput with the increase of the size
Varying Durations

But, in fact

Anticipate
Varying Probability Thresholds

[Bar chart showing throughput vs. threshold probabilities with legend for different strategies: naive, deadline-based, lazy, many-first, next-deadline, avg-deadline, threshold-based.]

Failed
Conclusions and Future Work

Conclusions

- Propose a new personalized anonymization method for LBSs considering not only locations but also the attribute observability.
- Propose several variations of strategies to implement the new anonymization method.
- Conduct experiments to evaluate the strategies.

Future work

- Develop high-throughput strategies that can anonymize users with low thresholds.
Thank you!

Have a nice day!