Preserving Location Privacy on the Release of Large-scale Mobility Data

Xueheng Hu, Aaron D. Striegel
Department of Computer Science and Engineering
University of Notre Dame
Introduction

Mobility models for wireless networks simulation

Synthetic models:
- Easy to set up
- Inadequate semantics to capture reality

Traces:
- Observed in real life, expected to be accurate
- Expensive to collect, privacy concerns

Challenge: trace publishing vs. location privacy
Preview

Trade-off: data utility vs. user privacy

Simulation scenarios
  DTN, Opportunistic N/W, etc.

Focusing on *realistic* wireless interactions - utility

Freeing from absolute locations - privacy
Computational Efforts

• Anonymization
  Creating ambiguity, e.g. k-anonymity (Sweeney, 2002)
Computational Efforts

• Obfuscation
  Degrading data quality (Krumm, 2009)

(a) Original GPS data
(b) Adding Gaussian noise
(c) Discretizing data to a grid
Data Source - NetSense Study

Provided 200 smart devices to incoming freshmen at Notre Dame (Aug 2011)

Sprint

User level agent (1-3 min polling intervals)

200x Nexus S 4G
200 anytime mins
Unlimited data, text

Environment (WiFi, Cell)
User proximity (Bluetooth)
Phone state (Screen, Battery)
Phone usage (data / app tonnage)

802.11n, 802.11g
Notations

- $MN = \{MN_1, MN_2, MN_3\}$
- $AP = \{AP_1, AP_2, AP_3\}$
- $R_L$: WiFi radio range
- $R_S$: B/T radio range
- $NL_i^t$ (WiFi Proximity Set), e.g., $NL_1^t = \{AP_1, AP_2, AP_3\}$
- $NS_i^t$ (B/T Proximity Set), e.g., $NS_i^t = \{MN_2, MN_3\}$

The vision of $MN_1$ at time $t$
Approach

Problem Description

Input: wireless relationships (WiFi and B/T proximity) constrained by $R_L$ and $R_S$
Output: trace solution for each mobile node preserving $\alpha$ of the given wireless relationships

Algorithm Components

Step 1: Access Point Deployment (2D)
Step 2: Mobility Trace Generation
AP Deployment

Connectivity Graph

- Connected if detected by the same mobile device
- Infer AP-to-AP relative distances (shortest path)
AP Deployment

Classic Multidimensional Scaling (Torgerson, 1952)

<table>
<thead>
<tr>
<th></th>
<th>BOS</th>
<th>CHI</th>
<th>DC</th>
<th>DEN</th>
<th>LA</th>
<th>MIA</th>
<th>NY</th>
<th>SEA</th>
<th>SF</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOS</td>
<td>0</td>
<td>963</td>
<td>429</td>
<td>1949</td>
<td>2979</td>
<td>1504</td>
<td>206</td>
<td>2976</td>
<td>3095</td>
</tr>
<tr>
<td>CHI</td>
<td>963</td>
<td>0</td>
<td>671</td>
<td>996</td>
<td>2054</td>
<td>1329</td>
<td>802</td>
<td>2013</td>
<td>2142</td>
</tr>
<tr>
<td>DC</td>
<td>429</td>
<td>671</td>
<td>0</td>
<td>1616</td>
<td>2631</td>
<td>1075</td>
<td>233</td>
<td>2664</td>
<td>2799</td>
</tr>
<tr>
<td>DEN</td>
<td>1949</td>
<td>996</td>
<td>1616</td>
<td>0</td>
<td>1059</td>
<td>2037</td>
<td>1771</td>
<td>1307</td>
<td>1235</td>
</tr>
<tr>
<td>LA</td>
<td>2979</td>
<td>2054</td>
<td>2631</td>
<td>1059</td>
<td>0</td>
<td>2687</td>
<td>2766</td>
<td>1131</td>
<td>379</td>
</tr>
<tr>
<td>MIA</td>
<td>1504</td>
<td>1329</td>
<td>1075</td>
<td>2037</td>
<td>2687</td>
<td>0</td>
<td>1308</td>
<td>3273</td>
<td>3053</td>
</tr>
<tr>
<td>NY</td>
<td>206</td>
<td>802</td>
<td>233</td>
<td>1771</td>
<td>2766</td>
<td>1308</td>
<td>0</td>
<td>2815</td>
<td>2934</td>
</tr>
<tr>
<td>SEA</td>
<td>2976</td>
<td>2013</td>
<td>2684</td>
<td>1307</td>
<td>1131</td>
<td>3273</td>
<td>2815</td>
<td>0</td>
<td>808</td>
</tr>
<tr>
<td>SF</td>
<td>3095</td>
<td>2142</td>
<td>2799</td>
<td>1235</td>
<td>379</td>
<td>3053</td>
<td>2934</td>
<td>808</td>
<td>0</td>
</tr>
</tbody>
</table>

(Source: https://personality-project.org/r/mds.html)
Trace Generation

AP deployment as the input

Location of a mobile device depends on locations of detected APs and MNs

Multiple solutions provide ambiguity
p1, p2, p3 – candidate solutions for mobile node x
Evaluation Metrics

Bluetooth proximity preservation w.r.t. $R_S$

- Precision: $P_{NS}$
- Recall: $R_{NS}$
- F_Score: $F_{NS}$

\[
P_{NS} = \left( \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{\text{original}_i \cap \text{retrieved}_i}{|\text{retrieved}_i|} \right) / (N \cdot T)
\]

\[
R_{NS} = \left( \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{\text{original}_i \cap \text{retrieved}_i}{|\text{original}_i|} \right) / (N \cdot T)
\]

\[
F_{NS} = 2 \cdot \frac{P_{NS} \cdot R_{NS}}{P_{NS} + R_{NS}}
\]

WiFi proximity preservation w.r.t. $R_L$ (similar)
Results

Bluetooth Proximity Preservation - Aggregate
Results

WiFi Proximity Preservation - Aggregate
Results

F_Score Distribution: Bluetooth vs. WiFi proximity preservation

**WiFi Proximity:**
~80% nodes ≥ 0.78

**B/T Proximity:**
~95% nodes ≥ 0.90
Results

Inter-contact Time Distribution:
Original samples vs. Generated traces

[Graph showing cumulative distribution of inter-contact time with two lines representing original data and generated traces.]
Summary

A novel approach to preserve location privacy using zero knowledge of actual locations

Potential impact where traces are needed but not available

Metrics to evaluate the preservation of wireless relationships
Open Problems

Introducing RSSI for better hint to relative distances

3D space deployment

Security challenges (quantitatively measuring how much privacy gained)
Questions?

Contact Information
Prof. Aaron Striegel : striegel@nd.edu
Xueheng Hu: xhu2@nd.edu

NetSense Study: http://netsense.nd.edu/