We Know Where You Live:
Privacy Characterization of Foursquare Behavior

Tatiana Pontes, Marisa Vasconcelos, Jussara Almeida, Ponnurangam Kumaraguru, Virgílio Almeida

Universidade Federal de Minas Gerais, Brazil
Indraprastha Institute of Information Technology, India

{tpontes,marisav,jussara,virgilio}@dcc.ufmg.br
pk@iiitd.ac.in

Pittsburgh, PA, USA - September 8th, 2012
Motivation and Goal

Great popularity of OSNs due to the easiness at which users can communicate and share content at large scale.

LBSNs arise with an additional attraction: association of geographic information with the shared data.

Foursquare: the currently most popular LBSN

Easy availability of information about location raises several concerns about privacy violation

Home location inference
Foursquare

David L. Lawrence Convention Center

venue
Wenda, I'm in the D. L. Lawrence Convention Center.
Foursquare

I'm the center mayor.

David L. Lawrence Convention Center

venue

I'm the center mayor.

mayorship
Foursquare

The center is amazing!! =)

David L. Lawrence Convention Center

venue
Foursquare

I agree with Wally!! The center is amazing.

David L. Lawrence Convention Center

venue

done
Foursquare

It's the 1° convention center I visit.

David L. Lawrence Convention Center

venue

badge
Foursquare

The center is amazing!! =)

It's the 1° convention center I visit.

I agree with Wally!! The center is amazing.

I'm the center mayor.

Wenda, I'm in the D. L. Lawrence Convention Center.

I'm the center mayor.

Wenda, I'm in the D. L. Lawrence Convention Center.
Foursquare

I agree with Wally!!
The center is amazing.

I'm the center mayor.

The center is amazing!! =)

done

tip

Publicly available attributes with geographic information associated.
## Dataset

*Dataset collected using the Foursquare API.*

<table>
<thead>
<tr>
<th>Crawling Period</th>
<th>August-October/2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>13,570,060</td>
</tr>
<tr>
<td>Number of venues</td>
<td>15,898,484</td>
</tr>
<tr>
<td>Number of mayorships</td>
<td>15,149,981</td>
</tr>
<tr>
<td>Number of tips</td>
<td>10,618,411</td>
</tr>
<tr>
<td>Number of dones</td>
<td>9,989,325</td>
</tr>
</tbody>
</table>
Public Attributes
- User Footprints -

About 30% of all users (~4.2 million) have at least one of the attributes.
Location Information
- Validity -

**User**
- home city (optional)

**Venue**
- city (required)

User Home City
- total users: 13,570,063 -

Venue Location
- total number of venues: 15,898,484 -

* Yahoo! PlacesFinder (geo-coding API)
Temporal and Spatial Analyses

Tips and Dones

Time stamp associated
Temporal and Spatial Analyses

- **Tips and Dones**
  - Time stamp associated
- **Inter-activity Time**
- **Displacements**
- **Returning Probability**
Temporal and Spatial Analyses

- **Inter-activity Time**
  Time between consecutive tips/dones of a user

  On average, 50% users have inter-activity time of at least 20 days.

- **Displacements**

- **Returning Probability**
Temporal and Spatial Analyses

- Inter-activity Time

- **Displacements**
  Distance between venues visited in sequence by a user

  On average, 70% of users have displacements of at most 150Km.

- Returning Probability
Temporal and Spatial Analyses

✔ Inter-activity Time

✔ Displacements

✔ Returning Probability
Probability to return to the same venue

On average, 52% of users have returns 1 week distant.
Methodology

**Technique:** Majority Voting Scheme

**Level of Inference:** Home city (also state and country)

**Ground-truth:** User's home city attribute

**Eligible Users:** Users with valid geographic information as the ground-truth:

- **Class 0:** one activity
- **Class 1:** multiple activities with a predominant location
- **Class 2:** multiple activities without a predominant location

**Metric for Evaluation:** Model accuracy
## Experimental Evaluation
- Class Distribution -

<table>
<thead>
<tr>
<th>Models</th>
<th>Eligible</th>
<th>Class Distribution (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Class 0</td>
</tr>
<tr>
<td><strong>Mayorship</strong></td>
<td><strong>1,814,184</strong></td>
<td>40.08</td>
</tr>
<tr>
<td><strong>Tip</strong></td>
<td>1,589,430</td>
<td>45.62</td>
</tr>
<tr>
<td><strong>Done</strong></td>
<td>1,194,907</td>
<td>45.762</td>
</tr>
<tr>
<td><strong>Mayorship+Tip</strong></td>
<td>2,521,338</td>
<td>35.63</td>
</tr>
<tr>
<td><strong>Mayorship+Done</strong></td>
<td>2,309,900</td>
<td>35.72</td>
</tr>
<tr>
<td><strong>Tip+Done</strong></td>
<td>2,093,120</td>
<td>39.74</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td><strong>2,823,404</strong></td>
<td>33.29</td>
</tr>
</tbody>
</table>
## Experimental Evaluation
- Home City Inference -

<table>
<thead>
<tr>
<th>Features</th>
<th>Class 0</th>
<th>Class 1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mayorship</td>
<td>51.61</td>
<td>67.41</td>
<td>60.12</td>
</tr>
<tr>
<td>Tip</td>
<td>51.52</td>
<td>67.29</td>
<td>59.11</td>
</tr>
<tr>
<td>Done</td>
<td>50.09</td>
<td>61.74</td>
<td>55.89</td>
</tr>
<tr>
<td>Mayorship+Tip</td>
<td>51.57</td>
<td>66.24</td>
<td>60.31</td>
</tr>
<tr>
<td>Mayorship+Done</td>
<td>51.05</td>
<td>65.27</td>
<td>59.51</td>
</tr>
<tr>
<td>Tip+Done</td>
<td>51.18</td>
<td>64.16</td>
<td>58.38</td>
</tr>
<tr>
<td>All</td>
<td>51.46</td>
<td>64.86</td>
<td>59.85</td>
</tr>
</tbody>
</table>
Experimental Evaluation
- Home City Inference -

1,339,152 correct inferences

<table>
<thead>
<tr>
<th>Features</th>
<th>Class 0</th>
<th>Class 1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mayorship</td>
<td>51.61</td>
<td>67.41</td>
<td>60.12</td>
</tr>
<tr>
<td>Tip</td>
<td>51.52</td>
<td>67.29</td>
<td>59.11</td>
</tr>
<tr>
<td>Done</td>
<td>50.09</td>
<td>61.74</td>
<td>55.89</td>
</tr>
<tr>
<td>Mayorship+Tip</td>
<td>51.57</td>
<td>66.24</td>
<td>60.31</td>
</tr>
<tr>
<td>Mayorship+Done</td>
<td>51.05</td>
<td>65.27</td>
<td>59.51</td>
</tr>
<tr>
<td>Tip+Done</td>
<td>51.18</td>
<td>64.16</td>
<td>58.38</td>
</tr>
<tr>
<td>All</td>
<td>51.46</td>
<td>64.86</td>
<td>59.85</td>
</tr>
</tbody>
</table>

1,504,262 correct inferences
# Experimental Evaluation
- Home State and Country Inference -

<table>
<thead>
<tr>
<th>Features</th>
<th>State</th>
<th>Country</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class 0</td>
<td>Class 1</td>
<td>Total</td>
<td>Class 0</td>
<td>Class 1</td>
<td>Total</td>
</tr>
<tr>
<td>Mayorship</td>
<td>71.27</td>
<td>80.92</td>
<td>76.70</td>
<td>89.79</td>
<td>92.92</td>
<td>91.64</td>
</tr>
<tr>
<td>Tip</td>
<td>70.29</td>
<td>80.59</td>
<td>75.53</td>
<td>90.12</td>
<td>93.67</td>
<td>92.02</td>
</tr>
<tr>
<td>Done</td>
<td>70.16</td>
<td>78.38</td>
<td>74.41</td>
<td>89.12</td>
<td>92.38</td>
<td>90.87</td>
</tr>
<tr>
<td>Mayorship+Tip</td>
<td>70.21</td>
<td>80.27</td>
<td>76.39</td>
<td>89.71</td>
<td>93.13</td>
<td>91.89</td>
</tr>
<tr>
<td>Mayorship+Done</td>
<td>70.01</td>
<td>79.89</td>
<td>76.07</td>
<td>89.18</td>
<td>92.78</td>
<td>91.47</td>
</tr>
<tr>
<td>Tip+Done</td>
<td>69.76</td>
<td>79.28</td>
<td>75.23</td>
<td>89.52</td>
<td>93.04</td>
<td>91.62</td>
</tr>
<tr>
<td>All</td>
<td>69.74</td>
<td>79.53</td>
<td>76.02</td>
<td>89.29</td>
<td>92.89</td>
<td>91.67</td>
</tr>
</tbody>
</table>

1,948,851 correct inferences

2,549,177 correct inferences
Understanding Errors
- Home City Inference -

- 46% of distances under 50km
- Greater distances suggest:
  - Travels
  - Inter-state mobility
  - Low user activity
Understanding Errors
- Home City Inference -

46% of distances under 50km

Greater distances suggest:
- Travels
- Inter-state mobility
- Low user activity

Home city inference within 50Km with around 78% of accuracy.
Conclusions

- Privacy violation and home location inference in a large Foursquare dataset.
- High accuracy in inferring where users live using publicly available attributes.
  - 60% for city, 77% for state and 92% for country inference.
  - 78% for city inference, with 50 km of error.
- Mispredictions may highlight some implicit user behavior in terms of mobility.

As future work:
- More sophisticated approaches.
- Finer granularity inferences (e.g., residence).
Thank you!
## Location Information

- Quality -

<table>
<thead>
<tr>
<th>Quality</th>
<th>% Users</th>
<th>% Venues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continent</td>
<td>0.0008</td>
<td>0.0007</td>
</tr>
<tr>
<td>Country</td>
<td>4.66</td>
<td>3.34</td>
</tr>
<tr>
<td>State</td>
<td>3.02</td>
<td>1.06</td>
</tr>
<tr>
<td>County</td>
<td>1.94</td>
<td>3.13</td>
</tr>
<tr>
<td>City</td>
<td><strong>80.02</strong></td>
<td><strong>78.70</strong></td>
</tr>
<tr>
<td>Neighborhood</td>
<td>7.58</td>
<td>12.03</td>
</tr>
<tr>
<td>Area of Interest</td>
<td>0.21</td>
<td>0.54</td>
</tr>
<tr>
<td>Street</td>
<td>2.53</td>
<td>1.08</td>
</tr>
<tr>
<td>Point of Interest</td>
<td>0.04</td>
<td>0.11</td>
</tr>
<tr>
<td>Coordinate</td>
<td>0.0005</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

Over 1.2 million users
Global Distribution

User Home City

Mayorship

Tip

Done
Global Distribution

User Home City

Mayorship

Denser map with high concentration in America, Europe and Southeast Asia

Tip

Done
Global Distribution

User Home City

Not all places that have users present venues

Mayorship

Tip

Done
Global Distribution

User Home City

Mayorship

Tip

Done

Sparseser map indicating that some cities have venues (mayors), but no tips.
Global Distribution

Even sparser with most activity concentrated in touristic/developed areas

Tip
Attribute Characterization

Top-3 cities with the highest number of:

- **Mayorships** - Jakarta, Bandung, Singapore.
- **Dones** - New York, Chicago, San Francisco.

Heavy tail distribution of attributes.
Temporal and Spatial Analyses
- Inter-Activity Time and Displacement -

![Graphs showing temporal and spatial differences](image)
Temporal and Spatial Analyses

- Returning Probability -

Half of returns concentrated in one hour mainly the first 10 min.