(Some) Social Aspects of Location Privacy

Kévin Huguenin (LAAS-CNRS / EPFL)
Séminaire Complex Networks, LIP6, 2015
About me

Researcher (2014-): TSF group, LAAS-CNRS
Postdoc (2011-2014): LCA1 group, EPFL
Postdoc (2011): DISL group, McGill
PhD (2007-2010): ASAP group, IRISA-Inria / Univ. Rennes 1
ENS Rennes

Other
- Nokia Research
- Telefonica I+D Barcelona
- Vrije Universiteit Amsterdam
Security of large-scale distributed systems

- Fault models
  - Byzantine faults
  - Crashes, losses

- Dissuasive methods
  - Decentralized mutual verifications
  - Collaborative detection and punishment

- Dishonest users
  - Rational people

- Hardware
  - Software
Privacy in mobile and social networks

• Mobile users, wireless networks, context-aware services, social networks

• Privacy issues

• Approach
  – Protection
    • Obfuscation, cryptography
    • Human and social aspects, privacy/utility trade-off
  – Quantification
    • Theoretical approach for privacy
    • Experimental approach for utility
  – Validation
Quantifying Interdependent Privacy Risks with Location Data

Alexandra-Mihaela Olteanu*, Kévin Huguenin*°, Reza Shokri*, Mathias Humbert*, and Jean-Pierre Hubaux*

*EPFL, °LAAS-CNRS, *UT Austin, ‡ETH Zurich

Location-based services are mainstream

- People leave *traces* with their location over time
  - infer location at unknown time moments
  - infer private attributes from whereabouts

- Location Privacy Protection Mechanisms (LPPM)
  - Hide location
  - Obfuscate location
Co-location is widespread

- **Declared** on online social networks
- **Inferable**
  - Pictures: face recognition
  - Mobile devices: behind NAT
  - Bluetooth: neighboring devices
Motivation

- Co-locations provide *information* and reduce user location privacy

- Co-locations add *complexity* to the inference process due to dependency among users
Formalization

• Extension of model proposed in [1]

\[ \text{Find nearby restaurants } @_{t=1} \{r_{26}, r_{27}, r_{35}, r_{36}\} \]

N users
M regions
T time instants

Adversary model

• Adversary: typically the service provider (e.g., OSN)

• **Prior knowledge** about users: Markovian mobility profiles [1]

• **Observations**
  – *Single* (obfuscated) observations \( u_{@t r} \)
  – *Co-location* observations \( u_i \leftrightarrow_t u_j \)

• **Goal**: infer the actual location of one or several users at some time instant
Co-location reporting function

\[ g_{u,v}^t(r, r') \triangleq \Pr(u \leftrightarrow_t v \mid a_u(t) = r, a_v(t) = r') \]

\[ g_{u,v}(r, r') = \begin{cases} 
\nu_{u,v} & \text{if } r = r' \\
\mu_{u,v} & \text{if } r \neq r'
\end{cases} \]

• Bluetooth scenario
• OSN scenario
  – Only for friends
  – Allow \textit{fake} co-locations
Optimal localization algorithm (FB)  

- Forward-backward algorithm with joint user variables \((r)\)

\[
\alpha^u_t(r) \triangleq \Pr \left( o(1) \ldots , o(t), C_1, \ldots , C_t, a(t) = r \mid \mathcal{K} \right)
\]

\[
\beta^u_t(r) \triangleq \Pr \left( o(t+1) \ldots , o(T), C_{t+1}, \ldots , C_T \mid a(t) = r, \mathcal{K} \right)
\]

- \(\alpha^u_t(r)\)  joint probability of all the observations up to and including time \(t\) and location of users at \(t\)

- \(\beta^u_t(r)\)  conditional probability of all observations after time \(t\), given the location of users at \(t\)
Optimal localization algorithm (FB)

- Forward-backward algorithm with joint user variables ($r$)

- Complexity:

\[
\begin{align*}
&\alpha^U_t(r) \triangleq \Pr(o(1), \ldots, o(t), C_1, \ldots, C_t, a(t) = r | K) \\
&\beta^U_t(r) \triangleq \Pr(o(t+1), \ldots, o(T), C_{t+1}, \ldots, C_T | a(t) = r, K)
\end{align*}
\]

- Localizing user $u_i$ at time $t$

\[
\Pr(a_{u_i}(t) = r) \{o(t)\}_{t=1..T}, C, K = \frac{\sum_{r \in \mathcal{R}^N} \alpha^U_t(r) \cdot \beta^U_t(r)}{\sum_{r \in \mathcal{R}^N} \alpha^U_t(r) \cdot \beta^U_t(r)}
\]

- Complexity: \((NTM^{2N})\)
Heuristic algorithm

• **Limited user-set approximation**
  – Perform the optimal algorithm on a *limited set of users* \((N=2/3)\)

\[
\text{co-target}_1(u) \triangleq \arg \max_{v \in \mathcal{U} \setminus \{u\}} |\{t \in \{1, \ldots, T\} | u \leftrightarrow_t v\}|
\]

\[
\text{co-target}_2(u) \triangleq \arg \max_{v \in \mathcal{U} \setminus \{u, u'\}} |\{t \in \{1, \ldots, T\} | u \leftrightarrow_t v\} + |\{t \in \{1, \ldots, T\} | u' \leftrightarrow_t v\}|
\]

– Complexity: \((NTM^4)/\ (NTM^6)\)
Approximation in Bayesian Network (BN)

- Loopy belief propagation (BP)
  - Approximate inference, iterative
Evaluation

- **GeoLife** GPS trajectory dataset by MSR Asia
- Synthetic co-location information
- Privacy metric: *Expected adversary’s error in localization*
Evaluation

• Individual LPPM for every user
  – **Location hiding** with probability $\lambda$:
    • do not report location information
  – **Location obfuscation** with probability $1-\lambda$:
    • report any of the actual location or it’s neighboring locations
Effects of co-location

- **Co-locations** bring significant *privacy loss*

\[ \nu = \text{Probability to report a true co-location} \]
\[ \mu = \text{Probability to report a fake co-location} \]

(b) With obfuscation

\[ \begin{align*} 
\nu = 0 & \quad \text{Light green} \\
\nu = 0.25 & \quad \text{Light green} \\
\nu = 0.50 & \quad \text{Green} \\
\nu = 0.75 & \quad \text{Dark green} \\
\nu = 1 & \quad \text{Dark green} 
\end{align*} \]

Privacy [km]

Location hiding probability (\(\lambda\))

Normalized privacy

Privacy loss [km]

Location hiding probability (\(\lambda\))
Comparison of inference algorithms

\[ n = 0.5 \]

\[ m = 0 \]

\( \gamma \) – Probability to report a true co-location

\( \mu \) – Probability to report a fake co-location

With obfuscation

\[ n = 0.5 \]

\[ m = 0 \]
Conclusions

• Propose generic framework to quantify location privacy loss stemming from co-locations

• Take-away message: others can compromise your privacy
  – A fast and efficient localization attack is possible
  – LPPM should not ignore the social aspects of location information

• Future work
  – Include social information to the adversary's knowledge
  – Make the framework more flexible
  – Design socially aware LPPM
Predicting Users’ Motivations Behind Location Check-Ins and Utility Implications of Privacy Protection Mechanisms

Kévin Huguenin°, Igor Bilogrevic†, Stefan Mihaila *, Joana Machado *, Reza Shokri ‣, and Jean-Pierre Hubaux*

*EPFL, °LAAS-CNRS, ‡UT Austin, †Google, ‣ETH Zurich

The Privacy / Utility Trade-off

• Utility is an overlooked, yet crucial, aspect of privacy protection
• Research effort
  – Large for privacy quantification and protection
  – Limited for utility (especially for subjective aspects, e.g., social networks)
  – Both are needed to design PETS
• Topic of this work
  – Focus on location check-ins *à la* Foursquare
  – Study the purpose of individual check-in
  – Study the effect of protection techniques on utility
Location Check-ins

- Check-ins at specific venues reveal **geographical** and **semantic** information (e.g., on Facebook or Foursquare)
  - Localization attacks can be improved by learning **patterns** at the **semantic** level
    - Evening: **restaurant** → **cinema**

- Protection? Obfuscation by **generalization**
  - Privacy? combination of geo and semantic patterns
  - Utility? Data-driven models based on personalized surveys

Venue—Food & Beverage—Restaurant—Burger Joint
Rue des Terreaux 10, 1003 Lausanne

I’m at **HolyCow!**

I’m in a **restaurant** in downtown Lausanne

After Holy Cow!, people often go to:

**KingSize Pub**
7.2 Rue du Port-Franç 16, Bar in Lausanne
# 1. Related Work

<table>
<thead>
<tr>
<th>Motivation for check-ins</th>
<th>Utility of check-ins</th>
<th>Location obfuscation</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Desire to connect with other and project interesting image of oneself</td>
<td>• Importance of audience of check-ins</td>
<td>• Well-studied topic in mobile networks</td>
</tr>
<tr>
<td>• Impression management</td>
<td>• Perception of check-ins by social circle</td>
<td>• Users lack awareness of long-term threats</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Limited effect on application functionality</td>
</tr>
</tbody>
</table>

- **Lack of user-centric utility functions for location check-ins**
  - Most prior works focus on the application dimension
    - (e.g., fraction of restaurants that are missed, error of traffic information, etc.)
  - We focus on the user, by predicting utility loss based on users’ perception
2. Methodology and Data Collection

- Personalized survey about Foursquare check-ins
  - Deployed over Mechanical Turk & ad-hoc Foursquare app
  - Provides ground-truth about location obfuscation through generalization

<table>
<thead>
<tr>
<th>Geographic information</th>
<th>Semantic information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Street number</td>
<td>Full semantic description</td>
</tr>
<tr>
<td>Street name</td>
<td>Venue type (1st ancestor)</td>
</tr>
<tr>
<td>City</td>
<td>Venue type (2nd ancestor)</td>
</tr>
<tr>
<td>Country</td>
<td></td>
</tr>
<tr>
<td>State</td>
<td></td>
</tr>
</tbody>
</table>

- Purpose of actual check-ins
- Utility of check-ins if sensitive location information is obfuscated
Original check-in

Purpose of check-in

Utility of check-in
3. Results

• Participants
  – 77 valid questionnaires
  • 43% male, avg. age 29 (± 6 y.), 96% from the US
  • 14% students, 12% education, 7% unemployed

• Purposes of Check-ins
  67% of all check-ins’ purposes related to a high-level social goal
Utility vs. Obfuscation Levels

• “On a scale from 1 to 5 (...), to what extent would your purpose be met, if
  the precise location 'would' be replaced with...

Ls = Low semantic obf.
Lg = Low geographic obf.
Hs = High semantic obf.
Hg = High geographic obf.

... an electronics store on E Dixon Blv, Shelby, NC”
...an electronics store in Shelby, NC”
... a shop and service venue on E Dixon Blv, Shelby, NC”
... a shop and service venue in Shelby, NC”
Utility vs. Obfuscation Levels

- Utility changes depending on the actual purpose of check-ins
  
  **Purpose: Inform about activity**
  
  - Utility 1
  - Utility 2
  - Utility 3
  - Utility 4
  - Utility 5

  **Purpose: Wish people to join me**
  
  - Utility 1
  - Utility 2
  - Utility 3
  - Utility 4
  - Utility 5

  - For **socially-oriented** goals, semantic obfuscation is worse than geographic obfuscation (across different purposes)
Inference of a Check-in’s Purpose

1. Raw check-in data

2. Feature engineering
   - Structured venue features
   - Unstructured text features
   - User features
   - Hybrid features

3. Inference and evaluation
   - Purpose inference
   - Correct classification rate
   - Obfuscation utility prediction
   - Utility as a function of check-in features, purpose and obfuscation levels

- Venue name, type
- # of check-ins
- Complete address
- Sentiment
- Emotion
- Ancestors in semantic hierarchy

- Purpose inference (WEKA)
  - SVM
  - Random forests
  - Logistic regression

- Utility prediction
  - Linear regression (R)
  - M5P tree (WEKA)
  - J48 classifier (WEKA)
Feature Engineering

• We only consider features that can be computed from the information provided by the LBSN service (user profile) and the check-in.

• Raw data → Features

- **User related features**
  - E.g.: User_ID, age_group, gender, home_city.

- **Location related features**
  - E.g.: Venue_city, venue_type, root2, root3.

- **Extra information about check-in**
  - E.g.: Has_collocations, part-of-day.

- **Text related features**
  - E.g.: Has_text, is_expression, is_event, nb_with, is_badge, has_mood, has_smile, has_invite_words, has_recommendation.

- **Hybrid features**
  - E.g.: Venue_name_text, venue_type_text, distance_to_hometown, same_city.
Purpose Inference

• 13-class purpose classifier

• In 60% of the cases, the actual purpose appears in the top-2 elements of the sorted list

• For 80% of them, it appears in the top-4 elements
**Purpose Inference: text vs. no text**

- To better understand how difficult it is to predict the motivation, split the test data in two sets: with or without text in the check-in.

<table>
<thead>
<tr>
<th>Check-ins without text</th>
<th>Random Forest CCR</th>
<th>Baseline CCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 motivations</td>
<td>51.8%</td>
<td>27.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Check-ins with text</th>
<th>Random Forest CCR</th>
<th>Baseline CCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 motivations</td>
<td>39.8%</td>
<td>26.3%</td>
</tr>
</tbody>
</table>

Check-ins without text are probably simpler
Modeling Utility vs. Obfuscation

• We can infer the purpose of a check-in
• How accurately can we estimate the utility of a check-in, after obfuscating it?

Results
– Linear model: $R^2 = 0.21$, mean error 1.18 over range [1,5] ($p < .01$)
  • Semantic obfuscation coefficient (-0.73) has a 82% more negative effect on utility as compared to geographic obfuscation (-0.4)
– Non-linear model (M5P model tree technique): mean error 0.66
– Cost-sensitive Classifier (J48): accuracy of 66.3%, mean error of 0.52
Summary and Future Work

• We propose an automated check-in purpose inference model, and evaluate loss of utility due to data obfuscation
• Purposes of check-ins mediate the perceived loss of utility due to obfuscation
  – Obfuscating check-ins’ data produces only limited effects on their perceived utility
    • For 60% of check-ins, some obfuscation causes no loss of utility
    • Semantic obfuscation is 2x worse than geographic obfuscation, in terms of utility for the users
• Possible to implement privacy-preserving features for location-sharing services, with minimal effect on usability
  – Propose by default optimal obfuscation level for given check-in
  – New purpose-specific features: “directions to venue” vs. “share picture”
• Future work
  – Explore differences across Location-Based Social Networks
  – Run a trial with a mobile application
Next steps: Privacy of life-style tracking systems

• Emergence of life-style tracking systems
  – Smartwatches, activity/sleep-trackers
  – Apple’s HealthKit / ResearchKit
  – Google’s AndroidWear

• Research directions
  – Inference of private information
  – User sharing attitudes, social aspects, and privacy perceptions
  – Privacy protection mechanisms
Next steps: Community-aware privacy attacks

• Information disclosed by individuals reveals information about people they are connected to

– Scalability issues for global inference
– Exploit communities to reduce scale
Effects of co-location

- **Co-locations** bring significant *privacy loss*

(a) Without obfuscation

- More co-locations → less privacy

\[ \nu = 0 \quad \nu = 0.25 \quad \nu = 0.50 \quad \nu = 0.75 \quad \nu = 1 \]

- Location hiding probability \( (\lambda) \)

\[ \mu = 0 \]
Effects of individual LPPM settings

Less privacy for a target user

More information available from her co-target

\( \mu \) – Probability to report a fake co-location

\( \upsilon \) – Probability to report a true co-location

Privacy loss induced by co-target

Self induced privacy loss
Effect of fake co-location information

\( l = 0.2 \)

\( \Downarrow \) – Probability to report a true co-location

\( \Uparrow \) – Probability to report a fake co-location

\( \nu \) – Probability to report a true co-location

\( \mu \) – Probability to report a fake co-location

With obfuscation

\( = 0.2 \)
Co-located users check-in at the same obfuscated location, if any.
Effect of user coordination

- Coordination improves privacy

Probability to report a true co-location: \( \nu \)
Probability to report a fake co-location: \( \mu \)

With obfuscation:
\[ \nu = 0.5, \quad \mu = 0 \]
Purpose Inference: hierarchical approach

Infer high level motivations: easier to classify, more balanced distribution: better results (CCR=64%)

Motivation
- Informative
  - Inform about location
    - Inform about people around me
    - Inform about activity
    - Keep track of the places I visit
- Utilitarian
  - Wish People to Join me
  - Recommend it
  - Appear cool/interesting
- Personal
  - Share mood
- Recreational
  - Get a reward
  - Participate in a game/competition
Purpose Inference: hierarchical approach

• Results for the classification:

Considering **13 motivations**:
- Baseline: CCR = 21.7%
- Random Forest: CCR = **45.3 %**

Considering **4 meta-motivations**:
- Baseline: CCR = 39.6%
- Random Forest: CCR = **64 %**

<table>
<thead>
<tr>
<th>Check-ins without text</th>
<th>Random Forest CCR</th>
<th>Baseline CCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 motivations</td>
<td>51.8%</td>
<td>27.5%</td>
</tr>
<tr>
<td>4 meta-motivations</td>
<td><strong>65%</strong></td>
<td>36.84%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Check-ins with text</th>
<th>Random Forest CCR</th>
<th>Baseline CCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 motivations</td>
<td>39.8%</td>
<td>26.3%</td>
</tr>
<tr>
<td>4 meta-motivations</td>
<td><strong>62%</strong></td>
<td>40.99%</td>
</tr>
</tbody>
</table>

Distribution of the hierarchical motivations for the check-ins with text:

- Game
- Personal
- Utilitarian
- Informative

Distribution of the hierarchical motivations for the check-ins without text:

- Game
- Personal
- Utilitarian
- Informative
Location privacy

• Emergence of location-based services

• Privacy issues:
  – Inference of private information (e.g., tracking)

• Protection mechanisms: obfuscation
  – Discretization, generalization, addition of noise
  – Pseudonymization

• Quantification
  – Inference using Hidden Markov Models with forward-backward algorithm
    (obfuscated) observation
  – De-anonymization using maximum likelihood matching

You are here
(and they know it)

Location privacy with co-locations

- Co-locations are widespread
- Co-locations provide *additional information* and further reduce users’ location privacy

- Co-locations add **complexity** due to dependency among users
  - Traditional solution: Hidden Markov Models
  - Problem: Users must be considered together; Solutions: small-scale, approx.
- Protection mechanisms: obfuscation
- Experimental validation w/ GeoLife mobility data set from MSR Asia
Location privacy with semantics

• Check-ins at specific venues reveal geographical and semantic information (e.g., on Facebook or Foursquare)

• Localization attacks are improved by learning patterns at the semantic level
  – Evening: restaurant → cinema

• Protection: Obfuscation by generalization
  – Privacy? combination of geo and semantic patterns
  – Utility? Data-driven models based on perso surveys
    • Amazon Mechanical Turk + Foursquare data
    • Median error of 0.51 (over 1-5) for utility

---


Location privacy with semantics

- Features:
  - userID, type of venue, has test, distance to hometown, ...
- Inference of motivation: 52%
- Inference of utility loss: median error of 0.52 (1-5 scale)
Mobile Information sharing

- Sharing information (or setting sharing policies) is usually a manual task
- Sharing decisions can be automated with machine Learning
  - Users request information
  - Decisions are made automatically based on contextual features
    - different granularities, e.g., street vs. city level
- Insights and validation through a personalized survey
  - Amazon Mech. Turk + Facebook data: What-if sharing scenarios
  - Statistical analysis (correlation, significance)
Mobile Information Sharing

- **Results**
  - Median number of **correct decisions** of up to 78% for automatic sharing decisions
  - Balance over-sharing (sharing when it shouldn’t) vs. under-sharing at limited cost

  ![Graphs showing correct decision proportions](image)

- **Implementation as an IM app over Jabber and on Android**


Location-based activity summaries

• Some activity-based online services (e.g., RunKeeper) rely on user-provided location traces
  – Location privacy and cheat-resilience issue
  – Other applications (health insurance, rewards)

• Solution for activity summaries
  – Lower-bounds of summaries
  – Based on a network of Wi-Fi access points

  – Find “best” samples: graphs-based formalization

Location-based activity summaries

- **Experimental validation w/ data sets from Garmin & FON**
  - Median accuracy of up to 81%
    (user runs 10km but gets a proof for 8.1km)

- **Evaluation of user awareness /concerns/acceptance through surveys**

  - Cheating opportunities (awareness: 0%)
  - Privacy issues (awareness: 19%)
Location-based activity summaries
HOT: Genomic privacy

- Increasing availability of genomic information
- Privacy risks (e.g., discrimination)
- Kin genomic privacy
  - Genomic information + family ties
  - Genomic 101
    - Reproduction
  - Inference based on loopy belief propagation on factor graphs
  - Evaluation: pedigree from Utah
- De-anonymizing genomic databases using phenotypic traits
