Unraveling an old cloak: k-anonymity for location privacy

Reza Shokri\textsuperscript{1}, Carmela Troncoso\textsuperscript{2}, Claudia Diaz\textsuperscript{2}, Julien Freudiger\textsuperscript{1}, and Jean-Pierre Hubaux\textsuperscript{1}

\textsuperscript{1} EPFL
\textsuperscript{2} K.U.Leuven
Location-Based Services

- Problems of revealing location to LBS provider
  - Sensitive information can be inferred from location
    - Example: being at the hospital, lawyer, church...
  - Profiling: daily routines, deviation from them, etc.
  - Stalking
  - Please rob me
  - ...

LBS Provider

query, user, location, time

local info, POIs, ...

Alice
Some location privacy techniques

- Mix zones [BS04]
- Location / time perturbation [HG07]
- Add dummy queries [CG09]
- Solutions based on $k$-anonymity
  - Anonymous usage of location-based services through spatial and temporal cloaking [Gruteser and Grunwald, 2003]
  - 650 citations in Google Scholar
  - Dozens of variants of the basic scheme have been proposed in the last years
k-anonymity in databases [SS98]

- Objective: ensure that an “anonymized” database record cannot be linked back to the data subject
  - Adversary has access to background info (e.g., census)
- Removing explicit identifiers (SSN, name, ...) not enough
- Quasi-identifiers: subsets of attributes that indirectly identify an individual (e.g., gender, DoB, zip code)
- k-anonymity: use generalization and suppression to ensure that each record could correspond to at least k individuals
## Example

### Release Table

<table>
<thead>
<tr>
<th>Race</th>
<th>Birth</th>
<th>Gender</th>
<th>ZIP</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>Black</td>
<td>1965</td>
<td>m</td>
<td>0214* short breath</td>
</tr>
<tr>
<td>t2</td>
<td>Black</td>
<td>1965</td>
<td>m</td>
<td>0214* chest pain</td>
</tr>
<tr>
<td>t3</td>
<td>Black</td>
<td>1965</td>
<td>f</td>
<td>0213* hypertension</td>
</tr>
<tr>
<td>t4</td>
<td>Black</td>
<td>1965</td>
<td>f</td>
<td>0213* hypertension</td>
</tr>
<tr>
<td>t5</td>
<td>Black</td>
<td>1964</td>
<td>f</td>
<td>0213* obesity</td>
</tr>
<tr>
<td>t6</td>
<td>Black</td>
<td>1964</td>
<td>f</td>
<td>0213* chest pain</td>
</tr>
<tr>
<td>t7</td>
<td>White</td>
<td>1964</td>
<td>m</td>
<td>0213* chest pain</td>
</tr>
<tr>
<td>t8</td>
<td>White</td>
<td>1964</td>
<td>m</td>
<td>0213* obesity</td>
</tr>
<tr>
<td>t9</td>
<td>White</td>
<td>1964</td>
<td>m</td>
<td>0213* short breath</td>
</tr>
<tr>
<td>t10</td>
<td>White</td>
<td>1967</td>
<td>m</td>
<td>0213* chest pain</td>
</tr>
<tr>
<td>t11</td>
<td>White</td>
<td>1967</td>
<td>m</td>
<td>0213* chest pain</td>
</tr>
</tbody>
</table>

### External Data Source

<table>
<thead>
<tr>
<th>Name</th>
<th>Birth</th>
<th>Gender</th>
<th>ZIP</th>
<th>Race</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andre</td>
<td>1964</td>
<td>m</td>
<td>02135</td>
<td>White</td>
</tr>
<tr>
<td>Beth</td>
<td>1964</td>
<td>f</td>
<td>55410</td>
<td>Black</td>
</tr>
<tr>
<td>Carol</td>
<td>1964</td>
<td>f</td>
<td>90210</td>
<td>White</td>
</tr>
<tr>
<td>Dan</td>
<td>1967</td>
<td>m</td>
<td>02174</td>
<td>White</td>
</tr>
<tr>
<td>Ellen</td>
<td>1968</td>
<td>f</td>
<td>02237</td>
<td>White</td>
</tr>
</tbody>
</table>

Figure 2: Example of k-anonymity, where $k=2$ and $Q=\{\text{Race, Birth, Gender, ZIP}\}$
**k-anonymity for “location privacy”**

- **Concept:**
  - Consider LBS query as an entry in the database
  - Consider location/time as a (quasi-)identifier
  - Generalize location/time so that it could correspond to \( k \) users
Privacy properties

- **Query anonymity**: not being able to identify the person who sent a query; i.e.:
  - not possible to link query to the user identity

- **Location privacy**: not being able to locate a user; i.e.:
  - not possible to link (location, time) to the user identity

- What do these k-anonymity techniques actually achieve??
Is $k$ related to location privacy?

- Ability to accurately locate the user related to the size of the region $R$, rather than to $k$
Adversary with access to user location in real time

- Query anonymity:
  - This adversary model and property are equivalent to what is achieved with k-anonymity in databases
  - Adversary uses background knowledge on the location of users to try to link users to their queries
  - k users are present in the region: queries are k-anonymous

- Location privacy:
  - Huh?
  - BY DEFINITION: Adversary knows the location of users!
Adversary with no background information

- Adversary only sees \(<q,R>\)
- Query anonymity: ANY user of the system could be the issuer of the query
- Location privacy: ANY user of the system could be in region \(R\)
- This has nothing to do with \(k\), or \(R\)
- With this adversary model, we do not need to construct any regions \(R\): just send \(<q,l,t>\) to the LBS and the same privacy is achieved.
Adversary with some (statistical) background information

- Query \( <q, R> \) is probably coming from one of the users likely to be in \( R \) (at the time the query is made).

- This is the adversary’s guess independently of how many users are actually in \( R \) when the query is sent!
  - Adversary cannot see who is actually there, so her guess is only based on \( <q, R> \) and the background information.
  - Whether there is actually 1 user or 3 users in \( R \), the query would be 3-anonymous towards the adversary.

- Constructing \( R \) based on who is actually in the region does not add anything.
In both cases, the query is 4-anonymous.

This technique actually enables the adversary to learn information she did not know!

- Adversary “expects” R (from background info), but instead receives R’
- Learns: A,B,E,F at home; C and D probably not at home
Conclusions

- No clear distinction in a large body of literature between “query anonymity” and “location privacy”
  - These are different properties!
  - k-anonymity techniques provide “query anonymity”, but not necessarily “location privacy” (as it is claimed)

- Location privacy does not seem to depend on k

- Constructing cloaking regions based on the actual location of k users leaks information to an adversary relying on statistical background info

- “Adapting” techniques from one type of systems (e.g., databases) to another (e.g., LBSs) is not straightforward

- Need to carefully define privacy goals and adversarial knowledge and strategy
  - This may sound obvious, but apparently it is not