Data Analysis from Mobile Networks

Petteri Nurmi
Questions

- What is the location area identifier? How it relates to location area code?
- What is a call detail record (CDR)? How it differs from data collected through mobile handsets?
- How CDR data can be used to estimate social networks?
- What kind of features can be extracted through CDR analysis?
- What kind of applications are there for CDR analysis?
Structure of Mobile Networks

- Mobile network divided into base stations and cells
- Base Transceiver Station (BTS):
  - Set of equipment that is responsible for handling communications within a specific geographic area
- Cell:
  - Geographic area covered by the same antenna
  - Each cell has an unique identifier (Cell ID)
- Each BTS responsible for one or more cells
  - Multiple directional antennas can be linked with the same BTS to increase channel capacity within the BTS area

Source: http://en.wikipedia.org/wiki/File:GSM_base_station_with_solar_panel_in_Sokosti_Finland.JPG
Mobile Networks

- Location Area
  - A cluster of cells referred to as a location area
  - Smallest unit for which the network maintains information about the location of a mobile unit
- Location Area Identifier (LAI)
  - Globally unique
  - Three parts: mobile country code (MCC), mobile network code (MNC) and location area code (LAC)
- Cell level information available to the network only when the terminal is active (i.e., transmitting)
Call Detail Record

- Record that is generated **by the network** when a phone call (or other type of connection) takes place
  - Making or receiving a phone call
  - Using a service (SMS/MMS/Data Connectivity)
- Contains (among other things):
  - Identifier of the initiating party (calling number)
  - Identifier of the receiving party (number or website)
  - Duration, type of connection, time
  - BTS responsible for handling the call
Call Detail Record

- Coordinates of the BTS can be used as a crude form of location estimate for the CDR
  - Coarse-grained location information, in densely populated urban areas accuracy typically within 1km
- CDR the most easily accessible source of data from the network side
  - Network also obtains information about handovers, details depend on specific network technology
  - Most common handover type change in the cell serving the handset, also channel change can trigger handover
CDR cf. Handset Data

- **CDR**
  - Coarse-grained location, BTS coordinates when handset active
  - Records generated when handset active (phone call, service usage)
  - Data can be collected from all handsets served by the network
  - Data collected on the network side and available to network provides
    Access to a larger population of users

- **Handset**
  - Identifier and signal strength of serving cell
  - Possibility to collect data continuously / at regular intervals
  - Data can be collected from selected handsets (requires software inst.)
  - Data collected on the client, available to person collecting data
    More accurate data and better control over it
CDR Data - Example
CDR Analysis

- Consumption analysis
  - User profiling based on communication patterns
- Social Network analysis
  - Detecting community/group structures, understanding true communication behavior, churn prediction etc.
- Mobility analysis
  - Characterizing mobility of an individual or a large population, estimating road or public transport usage
- Behavioral profiling
  - Socioeconomic or psychological profiling based on features extracted from CDR measurements
Example – Senseable City Rome

http://www.youtube.com/watch?v=RbhBz5UwRDQ
Social Network Analysis

- Social network is a theoretical construct used to study relationships between individuals (or groups)
- Defined by actors and ties
  - Actors (or nodes) represent the different individuals or groups
  - Ties (or edges, connections, links) represent relations between the actors
- Relationships can be
  - Two-way/Symmetric: directed graph
  - One-way/Asymmetric: undirected graph
From CDR to Social Network

• CDR, and other communication data, can be used to construct a multidimensional social network
  • Weighted graph, each edge linked with a vector of attributes
    – Number of calls made between two actors and their total duration
    – Message counts (SMS and MMS)
    – Calls during specific time (prime-time, weekend, etc.)
  • Weights represent the importance of the different attributes
    – Can be used to cluster the connections to subgroups (e.g., strong connections, sporadic connections, non-connections/noise etc.)
• Relationships can be constructed either considering symmetric or asymmetric connections
  • Symmetric: edge created when communication two-way
  • Asymmetric: edges created from all communication data
Degree, Geodetic Distance, Reciprocity

- **Degree**
  - Measures the number of connections that an actor has
  - In-degree: number of incoming connections
  - Out-degree: number of outgoing connections

- **Geodetic Distance**
  - Length of the shortest path connecting two actors

- **Reciprocity**
  - Actors A and B are reciprocal if there is a directed edge from A to B and vice versa
  - Degree of reciprocity: weight of the corresponding directed edges
Phone calls to and from number 8019

8019 -> 6171
4408 -> 8019
4002 -> 8019
8019 -> 4002
8019 -> 6171
4002 -> 8019
4408 -> 8019
8019 -> 4002

What are the in-degree and out-degree of node 8019?
2 and 2

What is the geodetic distance between 4002 and 6171?
2 (4002 -> 8019 -> 6171)

What is the degree of reciprocity between 4002 and 8019?
2
Centrality

- Measures the position of a node within a network
- Degree centrality
  - Number of adjacent nodes
- Closeness centrality
  - Sum of geodetic distances to all other nodes
- Betweenness centrality
  - Fraction of times a node lies along the shortest path between two other actors
- Eigenvector centrality
  - Connections to high-scoring nodes contribute more than connections to low-scoring nodes
  - Google’s PageRank is a variant of eigenvector centrality
Centrality - Example

<table>
<thead>
<tr>
<th>Node</th>
<th>8019</th>
<th>6171</th>
<th>4002</th>
<th>4008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree c.</td>
<td>2/2</td>
<td>1/0</td>
<td>1/1</td>
<td>0/1</td>
</tr>
<tr>
<td>Closeness c.</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Betweenness c.</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

6171 most central node in terms of in-links
4008 most central node in terms of out-links
Density, Cliques, Triads

- Density measures the overall connectedness of a social network
  - Defines as the number of connections divided by the total number of possible connections
- Clique
  - Subset of nodes that are connected by an edge
  - Clique is maximal if it cannot be extended by adding more nodes
- Triad
  - Structure consisting of three nodes and their mutual relationships
- Cliques and triads useful for identifying communities and subgroups
Density and Cliques - Example

Density of the network \(1/3\)
Density of the subgraph defined by node 8019 \(2/3\)

Only clique in the example is \(\{8019,4002\}\)

Triads: \(\{4408,8019,4002\}, \{4408,4002,6171\}, \ldots\)
Example –
Connected States of America

http://www.youtube.com/watch?v=eFKZnblAbol
Homophily Principle

- “Similarity breeds connection”
  - Personal networks tend to be homogenous with regard to many socio-demographic, behavioral and intrapersonal characteristics
- Selection
  - People tend to form relationships with others who are already similar to them
- Social influence
  - People adopt behaviors exhibited by persons they interact with
- Why it matters?
  - Churn: subscribers are more likely to transfer business elsewhere if they have close acquaintances that have transferred recently
  - Handset similarities: likelihood of having a particular handset increases with the number of close connections with the handset
  - Other user profile similarities can be detected as well
Mobility Patterns

• Individual mobility
  • Detecting important places
    – Home or work place detection
    – Other important locations
  • Degree of mobility
    – Area of influence and its diameter
    – Radius of gyration

• Collective mobility
  • Mathematical characterization of mobility
  • Estimation of traffic volumes
  • Estimation of origin-destination flows
Characterizing Mobility

- **Area of influence**
  - The geographical area within which a user spends most of her time doing daily activities
- **Diameter**
  - Maximum distance between two cell towers (BTS)
  - Characterizes the size of the area of influence
Radius of Gyration

- Radius of gyration
- Typical distance that an individual covers in daily mobility
- Calculation
  1. Assign a weight for each BTS observed in the CDR measurements
     - Relative usage of a BTS, e.g., in terms of number of records or total duration of traffic through a BTS
  2. Calculate a weighted center of mass using the BTS coordinates and weights
  3. Calculate the distance between each BTS and the weighted center of mass
  4. Radius of gyration is the mean of these distances
Radius of Gyration Example

<table>
<thead>
<tr>
<th>ID</th>
<th>Count</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>50417</td>
<td>1</td>
<td>60.1989</td>
<td>24.9529</td>
</tr>
<tr>
<td>40403</td>
<td>4</td>
<td>60.2063</td>
<td>24.9771</td>
</tr>
<tr>
<td>43651</td>
<td>2</td>
<td>60.2069</td>
<td>24.9482</td>
</tr>
<tr>
<td>40312</td>
<td>1</td>
<td>60.2040</td>
<td>24.9621</td>
</tr>
<tr>
<td>40332</td>
<td>33</td>
<td>60.2032</td>
<td>24.9655</td>
</tr>
<tr>
<td>40341</td>
<td>1</td>
<td>60.2094</td>
<td>24.9800</td>
</tr>
<tr>
<td>40342</td>
<td>264</td>
<td>60.2113</td>
<td>24.9729</td>
</tr>
<tr>
<td>40783</td>
<td>1</td>
<td>60.2089</td>
<td>24.9789</td>
</tr>
<tr>
<td>40801</td>
<td>1</td>
<td>60.2066</td>
<td>24.9796</td>
</tr>
</tbody>
</table>

Weighted centre of mass:
60.2103  24.9719

Distances:
1.64, 0.53, 1.37, 0.88, 0.86,
0.46, 0.13, 0.42, 0.59

Radius of gyration:
0.76km
Important Place Detection

- Basic idea same as in coordinate-based place detection
  - Detect geographic areas that contain numerous measurements within a spatially small area
  - Use temporal and spatial constraints to determine whether the place is meaningful or not
- CDR data cannot be reliably used to estimate the time the user spends in a location
- Importance can be estimated considering
  - Percentage of tower days, i.e., on how many days one of the BTS in a cluster was connected to
  - Number of days between first and last contact with any cell tower in the cluster
  - Number of days on which any BTS in a cluster was contacted compared to other clusters from a user
Important Place Detection - Example

```
50417  1  60.1989  24.9529
40403  4  60.2063  24.9771
43651  2  60.2069  24.9482
40312  1  60.2040  24.9621
40332  33 60.2032  24.9655
40341  1  60.2094  24.9800
40342 264 60.2113  24.9729
40783  1  60.2089  24.9789
40801  1  60.2066  24.9796
```

```
40403  4  60.2063  24.9771
40332  33 60.2032  24.9655
40342 264 60.2113  24.9729
```
Displacement $\Delta r$ defined as the distance between two successive CDR observations
- Corresponds thus to the distance between the cell towers that are associated with successive calls
- Distribution of displacements approximates a truncated power-law, also known as a truncated Levy-flight
- Accordingly, the distance that an individual travels follows approximately a truncated power law
- Central for estimating urban phenomena and for understanding behavior of individuals
- Similar characterizations have been obtained, e.g., for banknote dispersal

$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} \exp\left(-\frac{\Delta r}{\mu}\right)$$
Collective Mobility – Origin-Destination Flows

• Basic idea:
  • Detect start and end locations for trips using a place identification algorithm
    – Place identification necessary to overcome oscillations in location due to cell handovers
  • Divide geographic area under investigation into regions and assign a region for each origin and destination
  • Group trips with same origin and destination regions together using different temporal windows
    – Weekly, daily, hourly etc.
Socioeconomic Level

• Socioeconomic Level (SEL) (or status SES)
  • Characterization of an individual (or household) economic and social status relative to rest of the society
  • Typically defined as a combination of income related variables (salary, wealth, education)

• SEL Map
  • Characterization of socioeconomic level at different geographical areas
  • Typically indicator values stratified to a small number of levels, e.g., high, middle and low socioeconomic status
  • Computed by national statistical institutes
• Socioeconomic status correlates with features that can be extracted from CDR data
  • Behavioral features
    – Number of incoming/outgoing calls and messages, handset type, total duration of calls, expenses, …
  • Social network features
    – Degree of social network, frequency of communication, …
  • Mobility features
    – Diameter of the area of influence, radius of gyration, total distance traveled, …
• Socioeconomic status can thus be predicted by training a classifier/regression model
  • Features as before, class variable for training the SEL level of an individual as given by existing maps
Personality Detection

- The Big Five personality model
  - Five traits that characterize personality
    - Openness to experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism
  - Personality assessed using a questionnaire that presents positive and negative statements one at a time
  - For each trait, questionnaire returns a trait score that describes how characteristic the trait is in the individual
- Personality detection attempts to automatically determine personality traits
  - CDR data and social network data can be used to infer personality traits to some extent
Personality Detection

- Automatic personality detection can be seen as a regression problem
  - Estimate the score of a personality trait based on features extracted from CDR data and social network
- CDR variables:
  - Duration of received phone calls, number of received and placed phone calls, number of SMS/MMS sent/received etc.
- Social network variables:
  - Degree of the actor (number of contacts), density of subgraph formed by neighboring nodes
  - Number of strong contacts (frequent calling partners), efficiency to reach nodes
- Best results for openness, extroversion and agreeableness traits (mean squared error < .7)
Summary

http://www.youtube.com/watch?v=2aEPkyOBtRo
Summary

- Mobile networks
  - Divided into base stations and cells
- Call detail record
  - Record that is generated by the network when a call or other communication service is used
- CDR-based social network analysis
  - Communication patterns can be used to construct social network structure between people
  - Features such as degree, centrality, subgraph density etc. can be used to profile individuals
Summary

- CDR-based mobility analysis
  - Area of influence, diameter and radius of gyration features that characterize the spatial spread of an individual’s communication and movement patterns
  - Places can be identified analogously to the approaches used on coordinate data
- User modeling
  - Socioeconomic prediction
  - Personality detection
  - …


Literature

- Vieira, M. R.; Frias-Martinez, E.; Bakalov, P.; Frias-Martinez, V. & Tsotras, V. J. Querying Spatio-temporal Patterns in Mobile Phone-Call Databases *Proceedings of the 11th International Conference on Mobile Data Management (MDM), IEEE, 2010*, 239-248