Complex Network Analysis for Mobility Modeling - Practical Session

Prof. Thrasyvoulos Spyropoulos,
Pavlos Sermpezis

January 30, 2014

1 Lab Description

The dataset of a real network (opportunistic network experiment) is given and the goal is to analyze the data, understand their characteristics and calculate some important metrics.

A opportunistic network consists of mobile nodes which can exchange data using direct communications (e.g. Bluetooth, Wifi-direct, etc.), when they are in proximity, i.e. within transmission range of each other.

The dataset is from an experiment that took place in an academic conference on computer communications (*Infocom, 2006*). Devices with Bluetooth communication capabilities were given to some participants. Each device records the bluetooth sightings, i.e. when two users/nodes come within the Bluetooth transmission range and their devices manage to set up a connection, this event is recorded by the devices (*meeting or contact event*).

Since data exchanges take place only when nodes are in proximity (in contact), the time points, when the contact events take place, and the nodes involved, determine the dissemination of a message. Hence, we usually model such networks with the *Contact Graph* ("*CG*"). The *CG* is a graph (i) whose vertices represent the network nodes and an edge between two vertices implies that these two nodes can contact each other regularly; (ii) each edge between nodes ($\{i, j\}$) is associated with a random contact process with rate λ_{ij} (rate = contact frequency = number of contacts per time unit). These (random) contact times, define the times during which information can be exchanged between nodes $\{i, j\}$.

2 Dataset File

The dataset can be downloaded from the link:

- `http://www.eurecom.fr/~sermpezi/infocom.csv`

The csv-file **infocom.csv** contains information about the meetings between devices. Each line has the following format:

$$\{node1,\ node2,\ number\_of\_meetings\}$$

- *node1* and *node2* are the IDs (i.e. integer numbers starting from 1) of the nodes.
- *number_of_meetings* is the number of meetings (i.e. Bluetooth sightings) between *node1* and *node2*.

As an example, the first 5 rows are presented below:

```
1,    3,   75
1,    4,  114
1,    5,  210
1,   14,   92
1,   15,    1
```

3 Questions

1. Create a graph from the given file. (type of vertices: integers, weighted edges - weights: contact rates or, equivalently, number of meetings)
2. What is the number N of people participated in the experiment (i.e. number of nodes)? What is the percentage p of node pairs that have met each other? Would you characterize the network as *sparse* or *dense*?
3. What is the degree distribution of the graph? Find the (i) average, (ii) max, (iii) min degree and plot the CCDF of the degree distribution (the y-axis does not need to have probability values). Is the degree distribution power-law?.
   see example: `http://networkx.lanl.gov/examples/drawing/degree_histogram.html`
4. Plot the network graph. E.g. with **draw(...)** using the **spring_layout(...)** (you can use other layouts too).
   Plot the *betweenness centrality* (utility used both in SimBet and BubbleRap routing protocols) of different nodes (e.g. sorted values).
   If you wanted to route data based on a social utility (e.g. contact rates) in this opportunistic network, would the plots of the (i) contact graph, (ii) betweenness centralities, (iii) degrees be useful? Why?
5. Take into account only the (a) 50%, (b) 10%, (c) 5%, and (d) 1% of the *highest* rates (i.e. remove the edges with the lowest rates). As before, plot the (i) contact graphs (or the contact graph of the largest connected component), (ii) betweenness centralities, (iii) degrees, for all cases (a)-(d). Observe the differences between them. Which threshold would you use to infer information about the routing social utility? Why?
   hint: To be able to compare different cases simultaneously, for each case (a)-(d), make a copy of the initial graph G and then remove the edges. I.e. create 4 more graphs G50, G10, G5, G1.
6. What is the (i) average, (ii) max, (iii) min, number of times (or *rates*) that two nodes contact each other? Plot the CCDF of the rates in log-log plot. How are the contact rates distributed (i.e. which distribution(s) does the plot resembles)? Would you characterize the network *homogeneous* in terms of contact rates between its nodes and why?

Useful commands / methods / hints for NetworkX

- Read a graph as list of edges with numeric weights:
  read_weighted_edgelist(...)
- The command:
  degree()
  returns a *dictionary* with nodes as keys and degree as values or a number if a single node is specified. To get *only the values* of the degrees, use:
  degree().values()
  Similarly, the ending ".values()" can be used in other methods that return dictionaries.
- To get the edges and their weights, use:
  G.edges(data=True)
- If e is a weighted edge of a graph, its weight can be accessed by:
  e[2]['weight']
- To get the largest connected component as subgraph, use:
  connected_component_subgraphs(G)[0]