Project 1

EE232E - Graphs and Network Flows

In this project we will study a social network and graphs of users' personal friendship network. We will explore community structures in the friendship network and their interpretation and applications. All datasets are available online [1].

Submission: Please submit a zip file containing your codes and report to `ee232e.spring2016@gmail.com`. The zip file should be named as "Project1_UID1_UID2_..._UIDn.zip" where UIDx are student ID numbers of team members. If you had any questions you can send an email to the same address.

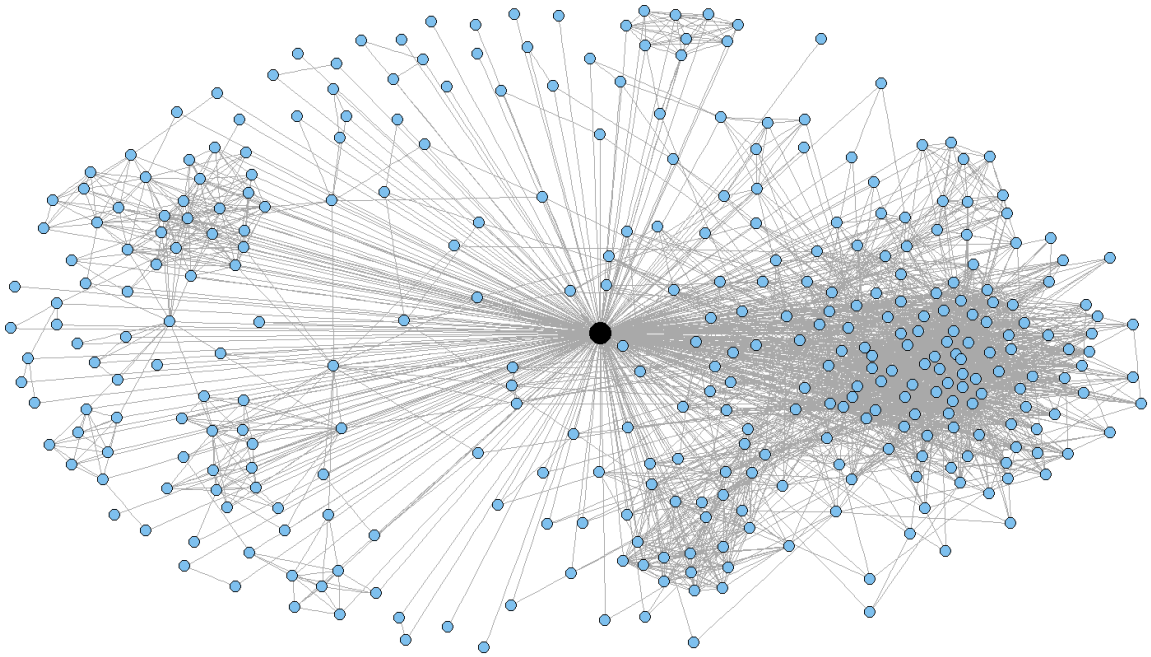
1. Download the Facebook graph edgelist file `facebook_combined.txt` [2] . Is the network connected? Measure the diameter of the network. Plot the degree distribution and try to fit a curve on it. What is your curve's total mean squared error? What is the average degree?

2. Take the first node in the graph (The node whose ID is 1) and find its neighbors. Create a graph that consists of node 1 and its neighbors and the edges that have both ends within this set of nodes. We will call this the personal network of node 1. Note that the common characteristic among the nodes in this graph -except for node 1- is that they are all friends of node 1, or equivalently, node 1 is a mutual friend of all of them. How many nodes and edges does this graph have?

3. Find nodes in the graph that have more than 200 neighbors (we will call these core nodes). How many core nodes do you find in the network? What is the average degree of these core nodes? For one of these nodes, find the community structure of the core's personal network. Plot the network and try to distinguish communities with color. Use Fast-Greedy, Edge-Betweenness, and Infomap community detection algorithms in igraph and compare results.

[1] `http://snap.stanford.edu/data`

[2] `http://snap.stanford.edu/data/egonets-Facebook.html`



An example of a personal network. The core node is shown in black.

4. Try removing the core node itself from its personal network and running the above community detection algorithms again. Are there any differences in the results?

5. Find $dispersion$ and $embeddedness$ for all nodes in the personal network. Embeddedness is the number of mutual friends a node shares with the core node and $dispersion$ is the sum of distances between every pair of the mutual friends a node shares with core node (These distances should be calculated in a modified graph where the said node and the core node are removed. In other words, between every pair of mutual friends, we consider the shortest path that does not pass through the said node and the core node. Read details in the paper referred [3]). Plot the distribution of $embeddedness$ and $dispersion$ over all personal networks created with core nodes from part 3. Plot 3 personal networks showing their community structure with colors and highlight the node with maximum $dispersion$ in each network. Highlight the edges incident to this node as well. On each network, do the same thing for the node with maximum $embeddedness$ and the node with maximum $\frac{dispersion}{embeddedness}$. Can you show and explain what characteristic of a node is revealed by each of these measures?

[3] "Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook", Lars Backstrom, Jon Kleinberg.
http://arxiv.org/abs/1310.6753

6. The communities in personal network can translate into different aspects of one's life. These can be friends from high school, college friends, colleagues, etc. Certain types of communities are present in almost every individual's personal network. Can you find structural features for each community that help you map communities across different people's personal network? That is, come up with a way that determines a community in one user's personal network is equivalent to another community in another users's personal network. You may not know the nature of the community in both networks (it could for example be "college friends" of each of the two users), but try to show similarities among both communities in the two networks that back up your decision. Run your algorithm across all personal networks you extracted and specify two types of communities you believe are recurring across all of them (along with the features you have calculated for each). Then, on each network you identify a community of type 1 and a community of type 2. You can only consider communities with size larger than 10 nodes from each personal network and ignore the rest. You may consider features like Modularity Index, Clustering Coefficient, Density, Community size, or any other statistical feature of the community.

7. Now we try to run the same kind of analysis on another real social network with tagged relationships. Download the Google+ ego networks file `gplus.tar.gz` [4]. Create personal network for users who have more than 2 circles (which is the default number). Extract the community structure of each personal network using both Walktrap and Infomap algorithms and show how communities overlap with the user's circles. How do these overlaps vary across users? How does this relate to a user's habit on tagging relationships with circles?

   Notes: Google+ ,unlike Facebook, has a directed network structure, where you can have someone in your circles regardless of whether they have you in their circles or they don't. Circles are tags you put on your relationships when you add people. E.g. you can have two circles, one named "friends" and the other one named "family", and when you add someone you can put them in one or both of these circles. You can refer to the dataset's readme file to see how it is organized. Notice that the ego node (core node) is not available in the edgelist stored for its personal network and you should manually add it.

[4] `http://snap.stanford.edu/data/egonets-Gplus.html`