STA 135 - Final Project

Clustering of Social Networks within Facebok

3-20-13 // Winter 2013 // Prof. Temple Lang

Hugh Crockford
Eliot Paisley
Evan Willenson

1 Introduction

This project originated with our group's collective curiosity in the analysis of social networks. Some of our originial ideas included predicting when two facebook users would be friends, creating a reccomendation system for users to find new friends, to exploring the internal structure of networks within facebook.

It was this last idea that we decided to go forward with for our analysis. The structure of this report is as follows: In Section 2 we present a detailed account of our process of acquiring the data we used, and then the code needed to turn it into a workable form. In Section 3 we discuss our research goals, and the motivation within. In Section 4 we present our results, and in Section some concluding remarks.

---

2 Data

Data Aquisition:

Our data comes from the Networking Group at UC Irvine.[1] Collected in 2009, "(the data is) A sample of 957K unique users obtained Facebook-wide by 28 independent Metropolis-Hastings random walks". Additionaly, we aquired a second file " ... contains additional node properties for each sampled user. For each sampled userID we have the number of times sampled, the total number of friends, the privacy settings and network membership." The sizes of these two files are 1.31GB, and 16.9MB, respectively.

With our goal of investigating the social networks within facebook, we set out to subset our data set based upon their network membership. Users are able to be members of multiple networks, which include regional, school, and workplace.

Go for it Hugh and Evan!

---

3 Research Goals

Our primary goal in investgating social networks within facebook is the ability to take what we've learned this quarter, and apply it to a real data set. As such, we have decided to perform a cluster analysis on ***600*** networks from our data set. In order to determine which networks are *similar* to each other, we first computed the following properties:

- Number of Nodes: A fundemental property of a graph, or network, the number of nodes for each of our networks with in facebook is simply the number of users belonging to each network.
- Degree-Total (DT): Another fundemental property, the *degree* of a node, is simply the number of edges that come out from the node. In our data set the degree of a user is simply the number of friends he or she has. To translate this into a property for the network as a whole, we compute the sum of the degrees over all users in the network. Thus, DT gives a measure as to how expansize the network is; small DT implies that the members of that network don't have as many friends as the members of a group with a high DT.
- Average Path Length (APL): defined as the "average number of steps along the shortest paths for all possible pairs of network nodes. It is a measure of the efficiency of information or mass transport on a network." In other words, given two users in a given network, we examine each user's list of friends to

[1] Minas Gjoka and Maciej Kurant and Carter T. Butts and Athina Markopoulou. *Walking in Facebook: A Case Study of Unbiased Sampling of OSNs.* Proceedings of IEEE INFOCOM '10, San Diego, CA. March 2010.

see if there is anyone in common. If so, we define the path length to be 1, and if not we examine each friend's list of friends to find a connection, and so on. Computing the APL for a network gives a sense to how closely related everyone is. Two versions of APL were calculated for our analysis; APL-C, and APL-U. The difference between these two measures was how we decided to deal with pairs of users in our network who *did not* have a connection after an exaustive search of the network.

APL-U treats these unconnected pairs as having the longest possible path-length; a value as long as the number of users in the network. APL-C, on the other hand, calculates the average path length as if the unconnected pairs weren't even there; they receive a path-length of zero. In a sense APL-U takes a conservative approach, treating everypone in the network equally, which APL-C is in essence measuring how closely related everyone in the network is, given that every member was already connected.

- Transitivity: While the calculations of APL, above, measures the distance to make a connection between two users, the *transitivity* of a users is the probability that the user's friends are ALSO friends themselves.
- Diameter: The diameter of a network is simply the longest path-length between two connected friends.

- Clusters: Within each network we can also examine how, if at all, the individual members cluster. Specifically, we calculate the maximal (weakly or strongly) connected components of a graph. Finding a (relatively) large number of clusters in a network implies that there may be many cliques, or subcommunities, while a small number indicates fewer distinct communities within the network.
- Graph Density: Graph density measures the ratio of the number of user pairs in the network that are connected divided by all *possible* user-user connections.

4 Results

test

5 Conclusion