Data Collection Report

Shailaja Singh, Xiaoni Cao, Justin Walker, Turn in (**a pdf**) through Canvas by 2:45pm:
Wednesday, February 21

1. How you obtained your data?
We obtained our data from Stanford Large Network Dataset Collection (SNAP) at https://snap.stanford.edu/data/egonets-Facebook.html.

2. How large is your data?
The Facebook dataset consists of 10 ego networks with 4039 nodes and 88234 edges from all 10 ego nets combined. Each of the 10 ego networks contains a circle network, an edgelist network, and a feature network (e.g., birthday, education, hometown, last name, first name, gender, languages, location, work etc.)

3. In what format are you storing your data. Describe the abstract data type, not just the file format.
Friendship networks can be modeled using an undirected graph where vertices represent people, and there is an undirected edge (v, u) if v and u are Facebook friends. In addition, we will use a m by n matrix to represent each ego network where each row is one of the nodes in the ego net, each column is one of the features in the ego net.

4. Did you need to process the original data to get it into an easier, more compressed format (e.g., convert from one format to another one)?
The data consists of two main parts, first a graph in which the vertices are your friends and edges are connections between your friends. Second part, the vertices (your friends) are further represented as vectors. These vectors are a superset of features representing you and your friends for eg. all the places where you and your friends have worked will get a column in this vector. To elaborate it further if your friend has worked at place X and this feature is captured at column 128 both you and your friend will have 1 at column 128 in this dataset. Since we are clustering the data w.r.t to the main user and not for the friends in the network we are reworking the vectors w.r.t the main user for eg. if your friend has worked at the same place as you, both of you will have 1 in the feature representing work 0 otherwise, if both of you have same hometown then both of you will have 1 for the feature representing hometown 0 otherwise. This way we can drastically reduce the features of the vector and at the same time ensure that the clusters formed are accurate enough.

5. How would you simulate similar data?
For this dataset the friends are represented as numbers for eg. if your network has 50 friends they are given unique numbers between 1-50. The main graph is given to us in the form of pairs of numbers representing an edge between friends. To simulate the main graph representing connection between friends we will be using a Python library NetworkX which is primarily used for network analysis but can also produce cluster-able graphs. For simulating the feature vector we are planning to use make_ blob feature in python to make cluster-able dataset.