Individual Assignment: 3

Predict 452 - Section 55 - Summer Quarter

Web and Network Data Science

Description:

For this assignment, I will be using the Facebook network data set to determine the most important individuals in the network and also will attempt to detect communities in the network, there are many ways to define the importance in the network, but in this assignment, i will consider betweenness centrality. The betweenness centrality is the measure of how many shortest paths pass through the particular vertex. The more the shortest paths that pass through the vertex, the more central is the vertex to the matrix. We will be using networkx python library for this assignment.

Data and implementation:

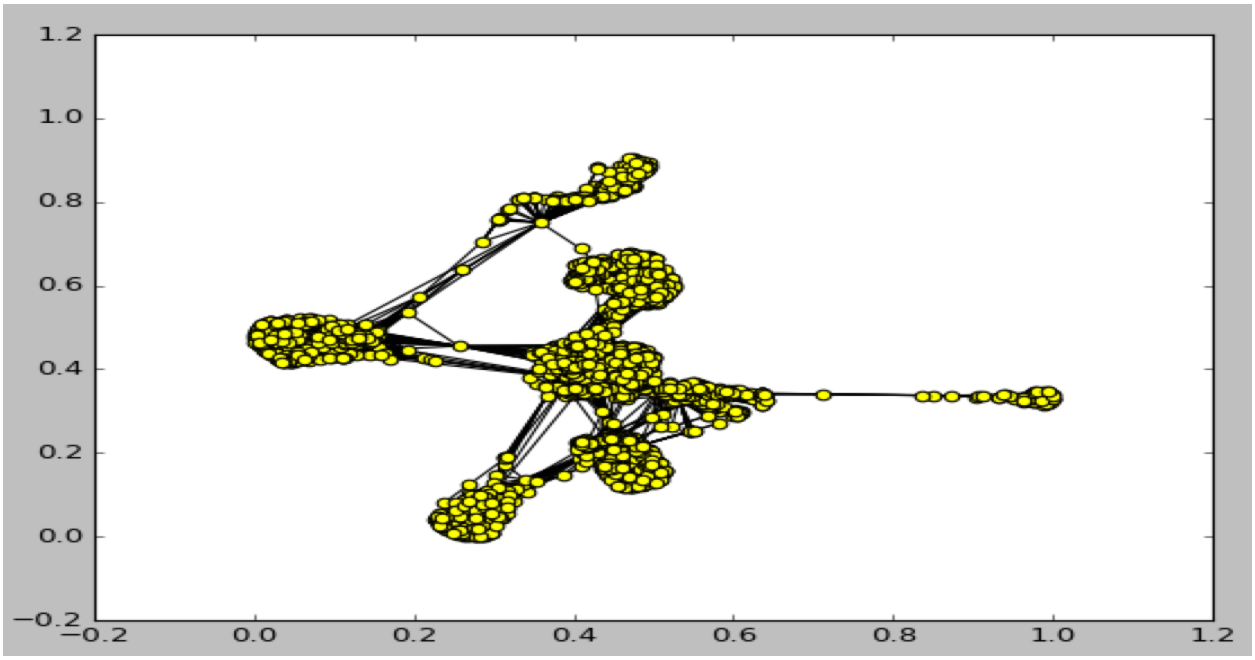
This dataset consists of 'circles' (or 'friend's lists') from Facebook. Facebook data was collected from survey participants using this Facebook app. The dataset includes node features (profiles), circles, and ego networks, Facebook data has been anonymized by replacing the Facebook-internal ids for each user with a new value.

Here's some basic information about the network dataset we will be working with:

Nodes: 4039
Edges: 88,234
Average Degree: 43.6910

This Facebook combined ego networks datasets contains the aggregated network of ten

individuals' Facebook friends list. In this dataset, the vertices represent individuals on Facebook, and an edge between two users means they are Facebook friends. Below is the Network.

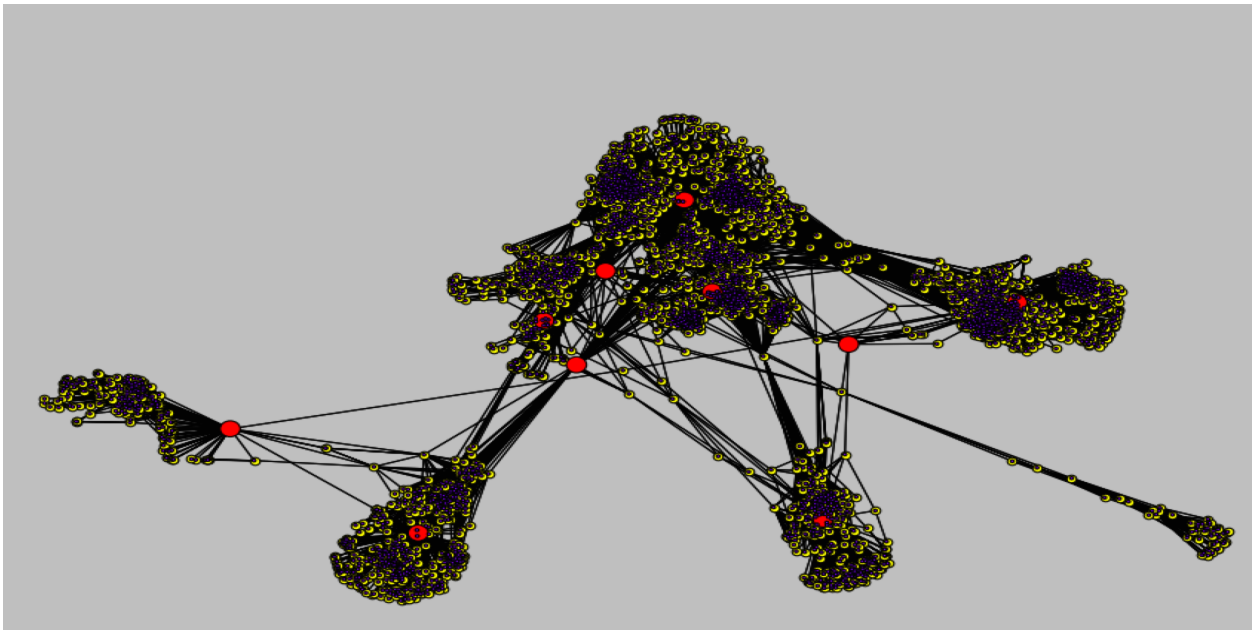


Fig(1)

As you can see from the above graph the network is densely clustered and we also notice the certain hubs of vertices appear.

Betweenness centrality:

To find the important individuals, i am using betweenness centrality as a measure, to do this I have used pool object from multiprocessing library and the itertools library, from the below graph we can see the 10 highest betweenness centrality measures (in red) in the network.



Fig(2)

as we can see, vertices that primarily either sit at the center of the hub or acts as a bridge between two hubs have higher betweenness centrality. The bridge vertices have high betweenness because all paths connecting the hubs pass through them, and the hub center vertices have high betweenness because all intra-hub paths pass through them.

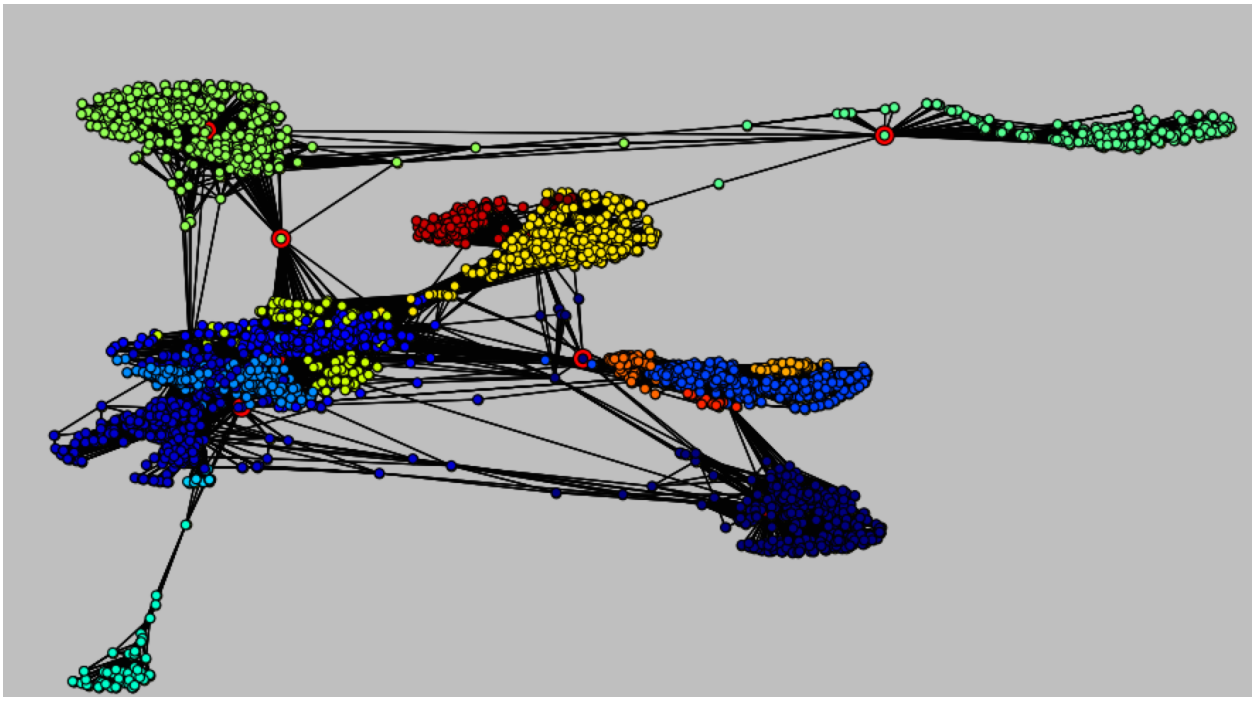
Community Detection:

Facebook friends come from different groups like school friends, work friends, friends from the hometown etc. In this assignment for the given data we trying to detect the communities and see if we can break down a social network into different potentially overlapping communities.

The criteria for finding good communities is similar to that for finding good clusters. We want to maximize intra-community edges while minimizing inter-community edges. I have used an open source library called community which is built on top of Networkx to detect the communities in the data. This algorithm tries to maximize the modularity of network, or the fraction of edges that fall within the community minus the expected fraction of edges if the edges were distributed by

random. Good communities should have a high number of intra-community edges, so by maximizing the modularity, we detect dense communities that have a high fraction of intra-community edges.

As you can see, the communities closely align with the vertex hubs. Because this dataset was compiled by aggregating the ego network of ten individuals, the different communities most likely reflect the different ego networks.



Fig(3)

Results:

From the above graphs, we can see that the, the vertices primarily sit at the center of the hub or acts as a bridge between two hubs have higher betweenness centrality, also for the communities the criteria for finding good communities is similar to that for finding good clusters. We want to maximize intra-community edges while minimizing inter-community edges. Formally, the algorithm tries to maximize the modularity of network, or the fraction of edges that fall within the community minus the expected fraction of edges if the edges were distributed by random.

Good communities should have a high number of intra-community edges, so by maximizing the modularity, we detect dense communities that have a high fraction of intra-community edges.

Reference:

Source citation

- J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.
- https://en.wikipedia.org/wiki/Betweenness_centrality
- https://app.dominodatalab.com/u/LeJit/FacebookNetwork/view/results/SocialNetwork.html
- https://blog.dominodatalab.com/social-network-analysis-with-networkx/