Seunghyuk Baek

[Sbaek44@vt.edu](mailto:Sbaek44@vt.edu)

CS5664 Spring 2018

Homework #2

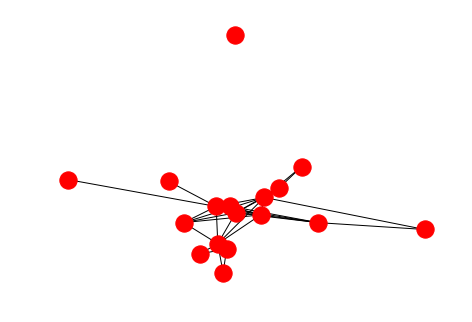
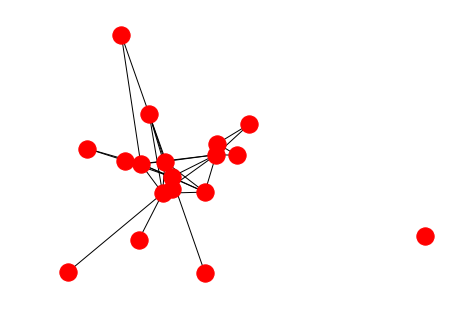
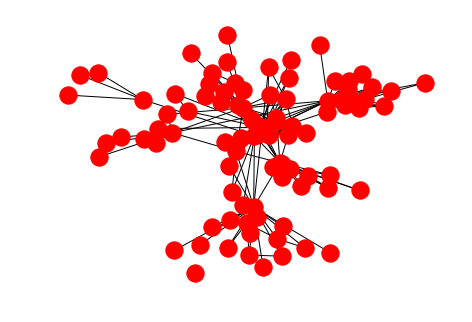
Community Analysis

For this analysis, the given dataset, Facebook data sample, was analyzed with Networkx library in Python and Gephi. First, I calculated modularity of the dataset with Gephi’s built in function. The result was 0.834. Since modularity is proportional, 0.8 mean this dataset can be divided into distinctive certain number of groups. To support this idea, we can calculate the average clustering coefficient for this network dataset. Both Gephi and Networkx result in 0.61 for the average clustering coefficient. This result means that there are some possibilities (higher than 0.5) which neighbors of a particular node are also connected each other. If we have a clustering coefficient that is very close to 0 or 1, it would be hard to estimate communities in the network since a coefficient close to 0 reflects hard to build groups and a coefficient close to 1 would mean every node is in one giant group. Now that we know the network has distinctive groups, modularity of top 3 communities of this network’s was calculated. The procedure is following. First, the network was dived into small communities by best\_partition function in community module. Second, found 3 large communities by sorting the number of nodes in each partition. Then modularity was calculated for those three communities. The modularity of entire network was 0.835. The largest community has modularity of 0.621. The second largest community has modularity of 0.403. The third largest community has modularity of 0.492. In fact, this network is composed of 17 communities. However, the modularity of top three community is not close to the modularity of the entire network. The result of this modularity is reasonable since communities in a network are already deviations from larger network. This also proves that the original network can be divided into several communities whereas communities of the original network is less likely to form small communities.

Next, K-clique analysis was performed to the largest community in the network. This can provide a vision of how each node is connected within a group. Since k-clique analysis for the entire network is time consuming, only the largest community was tested with k\_clique\_communities function in community module. In result, 17 was the largest k-clique in this community. There were two groups that satisfied 17 as k-clique number. The meaning of this K-clique analysis is that the largest community of the original network can be divided into 2 groups. To find how these communities are composed with smaller modules, average shortest path was calculated. The largest community has 1.97 average shortest path, the second largest community has 1.94 and the third largest community has 2.5. This means that first two communities have very similar structure while the third one has more scattered structure. For instance, the third community might have more linear structure compare to others.

Another way to partition a network is divide a network into its hierarchical communities. By using dendrogram, one can divide a network into three levels. The lowest level of dendrogram represent the lowest level of hierarchical tree. As level increases, each modules cluster together which eventually form the original network. First, clustering coefficients for these three partitions were calculated. The lowest level partition has clustering coefficient of 0.557, the mid-level partition has clustering coefficient of 0.520 and the highest cluster has clustering coefficient of 0.558. The modularity of those partitions were 0.814 for the lowest level partition, 0.835 for mid-level partition and 0.835 for the highest level partition. This is somewhat strange for me since I expected higher modularity or clustering coefficient as level increase, because higher level in dendrogram meaning each groups in lower groups together to form a community in higher level.

To Find the reason, the number of nodes in each level was calculated. At the lowest level, the number of nodes is 92 and the edge is 306. In mid-level of dendrogram, the induced graph has 18 nodes and 54 edges. The highest-level subgraph induced by dendrogram consists of 17 nodes and 52 edges. It seems nodes and edges can be grouped together to form larger groups proven by higher modularity value as dendrogram level gets bigger and fewer node and edge count at higher level dendrogram. The nodes in dendrogram represent communities not individual nodes in original network. I could conclude this by existing 17 nodes in level 2 dendrogram which is same as the number of groups in best\_partition function in community module.



Graph of induced subgraph from dendrogram level 0, 1, and 2

Next, for those partitions, k-clique analysis was performed. At lowest level, there were three subsets of 5-clique partitions (frozenset({0, 6, 10, 12, 18, 30, 31, 41, 42, 45, 47, 48, 49, 61, 73, 74, 76}), frozenset({2, 17, 18, 20, 21, 55}), frozenset({2, 18, 19, 34, 57})). At mid-level, there were two subsets of 5-clique partitions (frozenset({2, 6, 8, 12, 14}), frozenset({0, 4, 5, 6, 12, 14})). At last, at the highest-level there were 2 subsets of 5-clique partitions {0, 4, 5, 6, 12, 13}), frozenset({2, 6, 8, 12, 13}).

A close up of a map

Description generated with high confidence

This is the picture of the entire network generated by Gephi’s modularity ranking. Each color represents a partition. For this particular network we can see 6 partitions.

In conclusion, the original network is highly clustered and divided into several communities. there are 17 large communities in the original network. This can be estimated by modularity value of the entire network which is 0.8. Among those 17 communities, top 6 communities are the most recognizable. Furthermore, each community have different structure. For example, the largest and the second largest have smaller shortest path length compare to the third largest community.