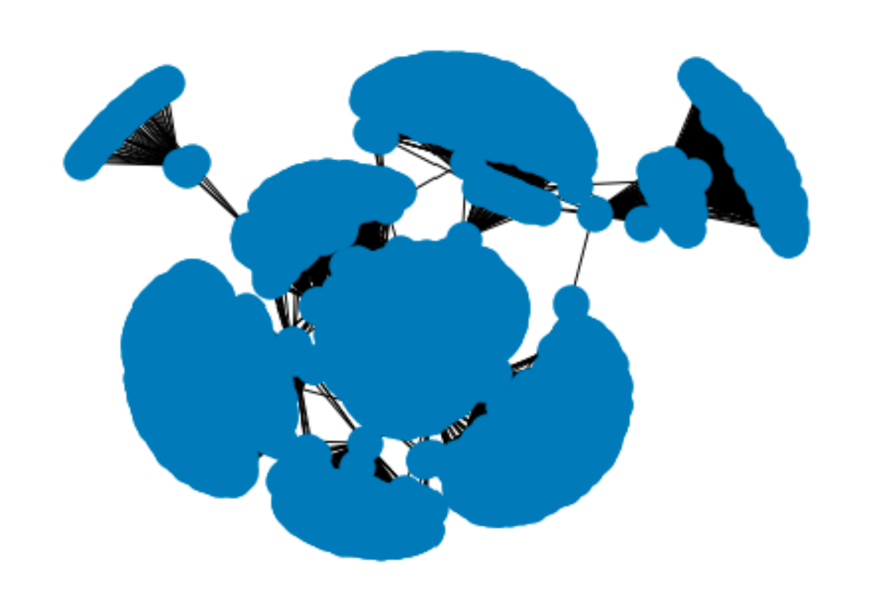
The goal of this project is to evaluate connecting community’s role of a central social network. A central node is generally an important person, a lead, an influencer, or head of a cult. We can study the question of if the relationship is one-sided or if the connections directly communicate with each other. Also, additional roles of a central node can be investigated based on the connecting nodes. For example, if a center node is a criminal, the surrounding nodes are generally criminals as well and they are connected around crime. In this project, we will study the Facebook dataset and how the nodes are connected to each other. The Facebook dataset forms a large network; therefore, it is computationally heavy to study. We will. Minimize the network size to be able to investigate it while maintaining the characteristics of the network. We will do this by visualizing the network and choosing a fraction of the network. Then, we will select the significant influencers of the network by using centrality metrics. We will apply clustering to find smaller networks as well. Further, we will run a graphical model to determine if we can create a community without influencer’s impact.

The Facebook dataset corresponds to a large network with 3663 nodes and 53498 edges. We can see from the following figure that the network forms various communities. These communities could be formed based on different features such as gender, nationality, interests, location, nationality, etc. However, we cannot determine based on the given dataset, which features are included in the study.



We can see from the figure that the network is very large and might be too large for a computer analysis even though we can visualize it. Therefore, we need to sample the dataset for the analysis. If we don’t sample, we may not be able to analyze the dataset using a simple computer.

One method of choosing a smaller subset is connected component. However, the results show that the network is only one connected component. Another method is to select one community to study. We will pick a dense cluster to get the most out of the experiment using a centrality metric: degree and betweenness. The results which are shown in a table demonstrate that 107 is the most connected node. Therefore, we will choose a network incorporating 107. The smaller subset has 333 nodes and 5038 edges.

The following table shows the top 10 highest scoring node based on degree centrality.

Table

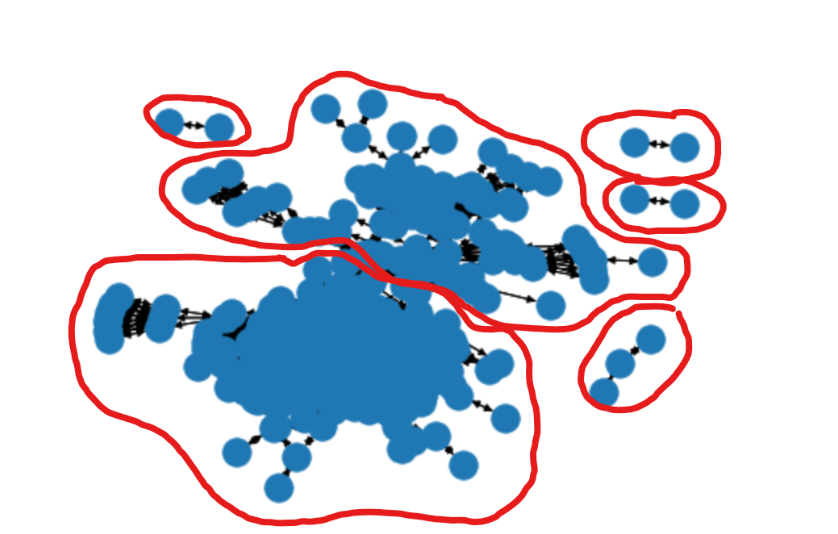
Description automatically generated

The following table shows the top 10 highest scoring node based on degree centrality

Table

Description automatically generated

The following figure shows the potential samples of the dataset. From these potential subnetworks, we will choose one containing 107.



Next, we will find the influencers in this network. Influencers are nodes that have impact on the connected nodes. Therefore, we will choose the betweenness centrality as a metric. We cannot choose metrics that are based on directed graph since we have an undirected graph in this study. Further, we cannot choose degree centrality due to its simplicity and closeness. Influencers are not necessarily close to their followers. In addition, influencers help create a community. Thus, we choose betweenness.

We choose top 1% betweenness centrality weight since their weights are greater than other nodes. These nodes have weights of 0.266 and 0.246 while most of the other nodes have a weight of less than 0.05. These nodes include: 277,175,19, and 23.

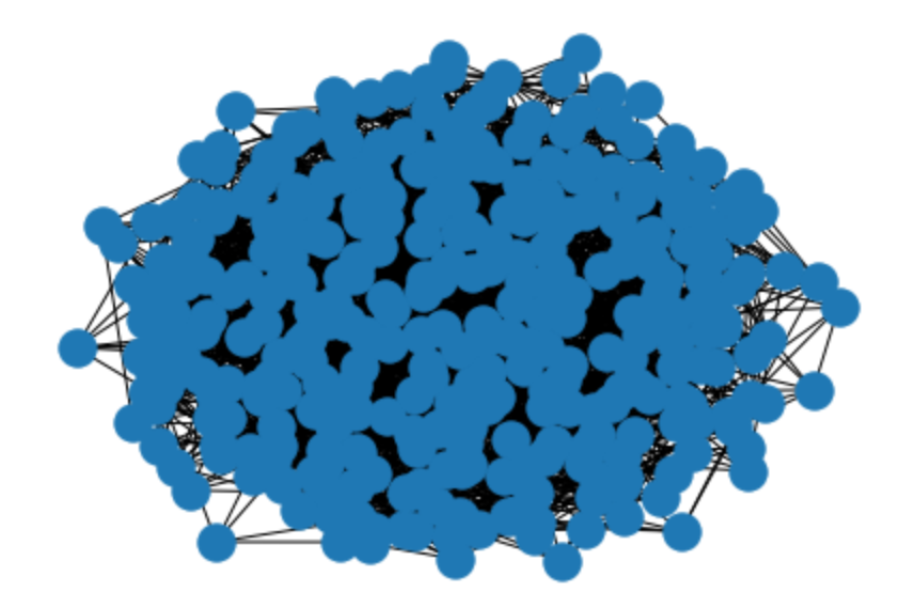
Further, we run a community detection algorithm with this new graph. The community detection algorithm detects subsets of nodes that are densely connected. The results are visualized in the following figure.

Chart, scatter chart

Description automatically generated

We can see from the plot that influencers had a major effect in engaging people in the Facebook network. Visually, a network without influencers is much sparser than the network with influencers. Although a visualization is simply not a great proof and a quantitative analytics method on a more complete dataset needs to be performed to sufficiently prove this conjecture. We cannot draw a better conclusion with the type of the data provided.

The studied sample has 333 nodes and 5038 edges which means that the number of possible edges are 110889. Therefore, the probability for a graph to occur is 0.045. Hence, if we use an Erdos-Renyi model, we can get a connected component graph as follows.



We can see from our study that we have some influencers in our network which was shown with the centrality weights. Without the influencers, we might have a major disconnection within the network. The information provided from the dataset makes it impossible to investigate further. However, further analysis is required to study the effect of influencers in our network. By comparing the two figures, we can conclude that the random generated model is connected but seem to be sparser than the actual network. This suggests that the users can be very connected inside a group. Also, there are some society parts who are separated from the rest of the society. Therefore, Erdos-Renyi is not generalizable to a human connection network.