**Exercise one: Analyzing an offline and online social networks**

Question 1 (3 points):

* Find out the node ID of
  + a) highest degree *Answer: S54*
  + b) highest betweenness *Answer:* S37
  + c) highest closeness *Answer:* S37
  + d) highest eigenvector in the Highschool network *Answer:* S110
* Highlight the above nodes in the Highschool network;

|  |  |
| --- | --- |
| Diagram  Description automatically generated | Chart, diagram  Description automatically generated |
|  |  |
| Chart, radar chart  Description automatically generated | |

* Explain why these metrics identify the same node or different nodes as the most central one.

Answer: So, the explanation may be given based on two arguments. First: the difference in their computations, and it provides the information about the difference between degree, and eigenvector centralities. With regard to similar “most-central” nodes for closeness and betweenness it is the fact that computations of both involve the shortest distances between the nodes.

What is more, we can think in terms of the interpretations of those centralities for degree, betweenness, closeness and eigenvectors which can be popularity of the node, brokerage of the node, the possibility to reach every other node, and the amount of influence of the node, respectively. We can see that in the networks of the real world the biggest popularity, for instance, does not imply the most influence, brokerage, or accessibility. Influence, on the other hand does not bring other concepts. However, the degree of brokerage is highly related to the ability to reach others, because a “node” would be chosen as broker only if the path through him/her is the shortest one.

Question 2 (5 points):

* Study the correlations between a) degree and betweenness, b) degree and closeness, c) degree and eigenvector *for all the nodes* in the Highschool network;

Answer: In Highschool data there is a strong correlation above 0,8 between degree and every other centrality. What should be noted is that for every pair of centralities with degree there is a threshold of the number of connected nodes at which the overall linear trend changes: around 5 nodes in the pair with closeness, at 10 nodes in the pair with betweenness, and around 12 in the pair with eigen centrality.

Chart, line chart

Description automatically generated

* Study the correlations between a) degree and betweenness, b) degree and closeness, c) degree and eigenvector *for all the nodes* in the Facebook network;

Answer: for this data which is 40 times bigger than the previous one, the correlation between degree centrality and others do not rise above 0.52 for the pair with betweenness, but gets higher above 0.7 for both eigen and closeness centralities. The metioned threshold is also present in that data: Around 250 nodes in the pair with closeness and in the pair with betweenness, and around 175 in the pair with eigen centrality.

Diagram

Description automatically generated

* From the above results, how well do different metrics correlate with each other? Which centrality metric will you use and why?

Answer: In the first data set the smallest correlations is between degree and eigen, but in the second case it is with betweeness. What is more, from the plots it is evident that closeness and degree receive of the biggest coefficient in both cases.

With regard to the question about what would we prefer to use, we would argue that it is hard to pick one because the most importand informaition is in their relations not in themselves. It also depends on which properties of the network and concepts related ot them we would like to study: popularity, brokerage, saturation, or the “influence ”. As we can see the correlations are different for both datasets, it could be because of the size or the real and virtual origins of the data, in our case is not that important, the important thing is that such conclusions as "your friends have more friends than you" [ J. Ugander et al.] can be derived only if we inspect those metrics together.

J. Ugander, B. Karrer, L. Backstrom, C. Marlow. The anatomy of the Facebook social graph.

Question 3 (5 points):

|  |  |
| --- | --- |
| Chart, bar chart  Description automatically generated | Chart, bar chart  Description automatically generated |
| Chart, scatter chart  Description automatically generated | Chart, histogram  Description automatically generated |

* For both the Highschool and Facebook networks, calculate the shortest path lengths between every pair of two nodes. How many percentage of nodes can be reached within 6 path lengths? Does “six degree of separation” apply to each network?

Answer: While in the dataset of High school there are more shorter paths (e.g. 1), the Facebook has an outstanding percentage of degree separation of 4; however for both dataset the six degree of separation works for the 90% of nodes. That shows that in this comparison the size of the network does not play that big role, only if we look at the number of other degree separations, but quantiles are similar.

* Study the degree distribution of these two networks, are they similar? Then use degree distribution to explain the degree of separation you answered above.

Answer: The degree distribution in case of Facebook is more skewed, but the prevalence of low connected nodes and an exponential reduction trend coming from less degree to higher degree are the same for both datasets. The six-separation degree phenomenon implies that even if some nodes have small number of connections, they can still be connected to anyone in the world through the other nodes to which they are connected. What is more the already mentioned paper of J. Ugander et al. is also related to these distributions: prevailing majority of the nodes populations is in the left part, and thus the overall probability of a random node to have more connections than other is really small, but the structure of networks and the presence of hubs(e.g. nodes from the first questions) make the “rule” of six degree possible by their brokerage properties.

Test the above hypothesis by the following steps (Question 4, 4 points):

1. Visualize the network and color the nodes by gender and residential hall, respectively.

Chart, radar chart

Description automatically generated

1. Build 8 subgraphs of the original network according to gender and residential hall: 1 subgraph for female student, 1 subgraph for male student, 1 subgraph for students with unknown gender, and 5 subgraphs for students living in residential hall from 1501 to 1505, respectively.

For example, to build a subgraph of all female students, you should keep all the nodes of female students and the edges between them. Other nodes and edges are removed

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| A picture containing diagram  Description automatically generated |

1. Study the edge density of all the subgraph and compare them to the edge density of the original network. What is your conclusion for the hypothesis?

Answer:

• Density for female friends is 0.0551786521935776

• Density for male friends is 0.0514285714285714

• Density for unknown friends is 0.1

• Density for 1501 friends is 0.12987012987013

• Density for 1502 friends is 0.0980392156862745

• Density for 1503 friends is 0.152046783625731

• Density for 1504 friends is 0.0758620689655173

• Density for 1505 friends is 0.0965909090909091

• Density for the whole network is 0.0536512667660209

Density computation is similar to the computation of the probability of a random graph: we divide the number of experienced outcomes by the number of the possible outcomes, which is completely the same with density formula (the number of present edges by the number of possible edges), so by comparing densities we also can compare the probability of a tie formation. Comparing the network density with all the subgraphs partly confirms the hypothesis for the study halls, as their densities are considerably different. However, for gender this is not the case only for the unknown gender, but these are small in size.

What is more, the simple comparison may be not sufficient for a statistical claim, so we suggest one of the possible methods (ERGM) and its output, which confirm our argument about the hypothesis: the significance of hall impact and insignificance of gender impact.

Call:

ergm(formula = net\_Highschool ~ edges + nodematch("gender") +

nodematch("hall"))

Maximum Likelihood Results:

Estimate Std. Error MCMC % z value Pr(>|z|)

edges -3.16066 0.08244 0 -38.337 <1e-04 \*\*\*

nodematch.gender 0.02325 0.10428 0 0.223 0.824

nodematch.hall 0.96893 0.10751 0 9.012 <1e-04 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null Deviance: 10232 on 7381 degrees of freedom

Residual Deviance: 3007 on 7378 degrees of freedom

AIC: 3013 BIC: 3034 (Smaller is better. MC Std. Err. = 0)

=Question 5 (4 points):

1. Calculate the modularity of the Highschool network if community is merely identified by a) gender and b) residential hall, respectively.
2. Search the Louvain Community Detection and explain the algorithm in your own words.
3. Use the Louvain Community Detection to identify communities in the Highschool network. Compare the modularity value produced by the Louvain algorithm to those in 1), and explain the reasons for the differences.

1. Calculate the modularity of the Highschool network if community is merely identified by

a) gender and

b) residential hall, respectively.

gender modularity 4.78267523722059e-05

hall modularity 0.17559113865932

2. Search the Louvain Community Detection and explain the algorithm in your own words.

The Louvain algorithm is an unsupervised community detection algorithmdivided in 2 phases: Modularity Optimization and Community Aggregation. The algorithm will start by randomly ordering all the nodes in the network in the modularity optimization phase. Then it will optimize modularity by merging communities of nodes until no significant increase in modularity is reached. After this phase all nodes belonging to the same community are merged in one big node to build a new network. In this network nodes represent communities from the previous phase and edges represent the sum of the weights of the edges between nodes in those communities. (Rita, 2020)

3. Use the Louvain Community Detection to identify communities in the Highschool network. Compare the modularity value produced by the Louvain algorithm to those in 1), and explain the reasons for the differences.

Modularity by the Louvain algorithm: 0.701644

The modularity of the Louvain algorithm is way higher compared to the networks identified by gender or residential hall. This is because the Louvain algorithm tries to create communities which improves the modularity.