Lab 1b: Descriptive Network Analysis – Local and Global Network Properties

CompSci 396-0: Social Networking Analysis ***Win 2022***

Student Name: Jiaqi Guo NetID: JGR9647

● Responses to Question

■ Part III: Individual Network Properties

1. **(10 points) Provide a table ranking the top 5 nodes in your network on each centrality measure. Each centrality means (a) in-degree, (b) out-degree, (c) betweenness, (d) in-closeness, (e) out-closeness, (f) eigenvector, (g) Burt's network constraint, (h) hub score, and (i) authority score**

Ranking	in-degree	out-degree	betweenness	in-closeness	out-closeness	eigenvector	net constraint	hub score	authority score
Top 1	59	1	1	149	131	1	164	1	59
Top 2	210	193	247	59	185	105	175	119	53
Top 3	149	247	19	23	160	224	199	57	105
Top 4	23	53	218	210	46	59	284	247	224
Top 5	1	19	193	182	226	53	91	251	70

The numbers in the table represent the indexes of nodes(identity). There are 290 nodes in this figure, corresponding to node indexes 1-290

2. **(10 points) Briefly describe each centrality measure. How is each computed and what does its number mean in your network (e.g., a high centrality score means…)?**

(a) in-degree

In-degree centrality measures the number of edges others have initiated with a vertex.

$$\mathrm{N}^{+(\mathrm{v})} = \{\mathrm{i} \in \mathrm{V(G)}: (\mathrm{i},\mathrm{v}) \in \mathrm{E(G)}\}$$

A high In-degree centrality value indicated other nodes in the network have a high level of engagement with a node.

(b) out-degree

Out-degree centrality counts the number of edges a vertex has initiated with others.

$$N^{-(v)} = \{i \in V(G): (i,v) \in E(G)\}$$

A high out-degree centrality value indicates a high level of engagement a node initiates with other nodes of the network community.

(c) betweenness

Betweenness centrality is a way of detecting the amount of influence a node has over the flow of information in a graph.

The betweenness of a vertex, v, is given by:

$$\mathrm{B(v)} \ = \ \mathrm{sum}\left(\frac{\mathrm{g_{ivj}}}{\mathrm{g_{ij}}}, \mathrm{i} \neq \mathrm{j}, \mathrm{i} \neq \mathrm{v}, \mathrm{j} \neq \mathrm{v}\right)$$

Where g_{ivj} is the number of geodesics from i to j through v. Therefore, a high-betweenness vertices usually lies on a large number of non-redundant shortest paths between other vertices. Which can be considered as “bridges” or “boundary spanners”.

(d) in-closeness

The closeness centrality of a vertex is defined by the inverse of the average length of the shortest paths to/from all the other vertices in the graph.

$$\frac{1}{\mathrm{sum(d(v,i), i \neq v)}}$$

The in-closeness can be considered as an index of the expected time-until-arrival for information in-flowing from other nodes to a certain node through the network via optimal paths.

(e) out-closeness

The closeness centrality of a vertex is defined by the inverse of the average length of the shortest paths to/from all the other vertices in the graph.

$$\frac{1}{\mathrm{sum(d(v,i), i \neq v)}}$$

The out-closeness can be considered as an index of the expected time-until-arrival for information out-flowing from a node to others through the network via optimal paths.

(f) eigenvector

Eigenvector centrality scores correspond to the values of the first eigenvector of the graph adjacency matrix.

In general, vertices with high eigenvector centralities are those which are connected to many other vertices which are, in turn, connected to many others.

(g) Burt’s network constraint

Burt’s network constraint is commonly used as a measure of structural holes.

Burt's measure of constraint $C[i]$, of vertex i's ego network $V[i]$, is defined for directed and valued graphs,

$$C[i] \ = \ sum(\,[sum(\,p[i,j] + \ p[i,q]p[q,j], q \,\epsilon\, V[i], q \neq \ i,j\,)]^2, j \in V[i], j \neq i)$$

for a graph of order N, where proportional tie strengths are defined as

$$p[i,j] = \frac{(a[i,j] + a[j,i])}{sum(a[i,k] + a[k,i],\ k \in V[i], k \neq i)}$$

$a[i,j]$ are elements of A and the latter being the graph adjacency matrix. For isolated vertices, constraint is undefined.

The larger the constraint value, the less structural opportunities a node have for bridging structural holes.

(h) hub score

Hubs and authorities are a natural generalization of eigenvector centrality.

Let A be the adjacency matrix of a directed graph. The hub centrality matrix R is given by:

$$R = A * A^T$$

A high hub actor points to many good authories and a high authority actor receives from many good hubs. The hub score is proportional to the authority scores of the vertices on the out-going ties.

(i) authority score

Hubs and authorities are a natural generalization of eigenvector centrality.

Let A be the adjacency matrix of a directed graph. The hub centrality matrix C is given by:

$$\mathrm{C} = \mathrm{A}^{\mathrm{T}} * \mathrm{A}$$

A high hub actor points to many good authories and a high authority actor receives from many good hubs. The authority score of a vertex is therefore proportional to the sum of the hub scores of the vertices on the in-coming ties.

3. (5 points) How does the centrality of nodes vary with different types of centrality metrics? Why is this the case? Please offer some potential explanations using certain nodes as examples

Example node 1: This node "[deleted]" has the highest out-degree, betweenness, eigenvector centrality and hub score. And its in-degree centrality is also the top 5 among the network community.

Node index (Name)	in-degree	out-degree	betweenness	in-closeness	out-closeness	eigenvector	net constraint	hub score	authority score
1	21	22	6378.9199134	1.604e-05	1. 896e-05	1.000000e+00	0.08152792	1.000000e+00	1.0221e-01

It is not difficult to find that the in-degree and out-degree centrality of this node makes it play a very important role in the network, which also means that this node can greatly influences the information transmission of network. Therefore, this node also owns a high betweenness centrality. The high overall degree centrality also means a high eigenvector centrality. In addition, as this node has a large out-degree, it is more likely to point out some authoritative information in the network, which makes it more likely to become a hub (a high hub actor points many good authories)

Example node 2: This node "Chickenman1964" has the highest in-degree centrality and authority score. And its in-closeness, eigenvector centrality is also the top 5 among the network community.

Node index (Name)	in-degree	out-degree	betweenness	in-closeness	out-closeness	eigenvector	net constraint	hub score	authority score
59	42	0	0	2.112e-05	1.193e-05	4.288355e-01	0.06452707	0.0000e+00	1.00000e+00

This node has the largest in-degree centrality, but its out-degree centrality is 0, that is to say, for this node, the information flow is one-way, which makes this node cannot affect the information transmission of rest parts of the network. In other words, the node's betweenness centrality is absolutely 0. And, based on the definition of authority score, the highest in-degree centrality makes it capable to receive information form many good hubs and give it the highest authority score. Besides, its high in-closeness centrality also enables it to receive information efficiently.

Part IV: Global Network Properties

1. (3 points) Briefly describe

(a) what k-core is:

The k-core of graph is a maximal subgraph in which each vertex has **at least degree k** (a node must have at least k links to other nodes in the k-core regardless of how many other nodes they are connected to outside the k-core). The cores of a graph form layers: the $(k+1)$-core is always a subgraph of the k-core.

(b) what insight this k-core decomposition method provides:

The k-core decomposition is to find the largest subgraph of a network, in which each node has at least neighbors in the subgraph. And it is usually the core component in the network.

(c) what is the highest/maximum level, k, of cores present in your network (e.g., Do any 3-cores exist in your network? Do any 4-cores? 5-cores? etc.)?

The highest level of k that present in my network is 4

2. (3 points) Visualize your network using k-core decomposition and include the visualization in your report. In a paragraph, discuss your interpretation of the visualization and whether the results of k-core decomposition make sense based on your expectations of the network.

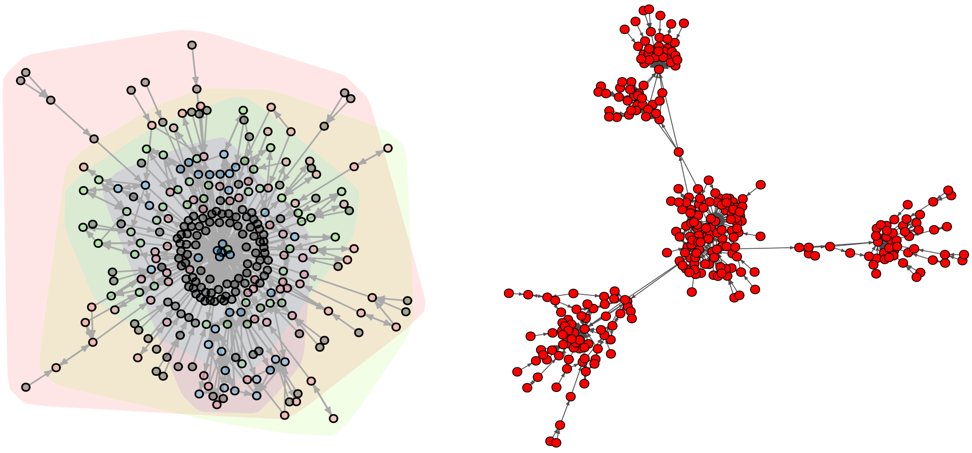


Figure 1 **Left:** the k-core decomposition **Right:** The visualization of the network

Based on the origin visualization of graph, we can easily observe that this network has 4 obvious nodes clusters which correspond to the 4 biggest colorful contours in the k-core decomposition visualization.

This makes sense because k-core is more likely to occur in these node clusters, and the number of clusters is exactly equal to the number of subgraphs in k-core Decomposition.

3. **(3 points) Pick one of community detection algorithms to run on your network. Which community detection algorithm did you choose and why?**

Detection Algorithms	Infomap community finding	Walktrap (short random walks)	edge betweenness
Modularity	**0.67**	**0.71**	**0.50**

Using **Walktrap**, we can obtain the following result:
IGRAPH clustering walktrap, groups: 32, mod: 0.71
This algorithm is chosen because it has the highest modularity score, 0.71, which suggests that this community division pattern has the strongest structural strength.

4. **(3 points) How many communities have been created? For your network, what might a community of nodes potentially have in common?**

32 communities have been created. For my network, all nodes in a community may have similar views or opinions toward my research topic——"San Jose Moves to Require Gun Owner to Have Insurance and Pay Annual Fees".

5. **(3 points) What is a modularity score? Interpret the modularity score of your results of community detection?**

The **modularity** of a graph with respect to some division (or vertex types) measures how good the division is, or how separated are the different vertex types from each other. Mathematically, modularity reflects the concentration of edges within modules compared with random distribution of links between all nodes regardless of modules, which can be defined as:

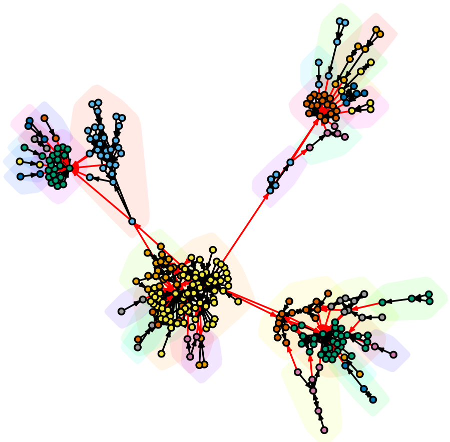
$$Q = \frac{1}{2} * \sum_{ij} [A_{ij} - \frac{k_i * k_j}{2m}] \delta(C_i, C_j)$$

here m is the number of edges, A_{ij} is the element of the A adjacency matrix in row i and column j, k_i is the degree of i, k_j is the degree of j, c_i is the type (or component) of i, c_j that of j, the sum goes over all i and j pairs of vertices, and $\delta(\mathrm{x,y})$ is 1 if $x = y$ and 0 otherwise.

Our modularity score **0.71** indicates the high concentration of edges within each divided community in our network.

6. **(3 points) Plot the communities and include the plot image in your report. What information does this layout convey? Are the communities well-separated, or is there a great deal of overlap? Describe the actors between any components and cliques (i.e., brokers). What are common features of these actors?**

It conveys how the network are formed by separate communities. All the communities are well-separated. These communities are usually connected by a node with high centrality. Besides the connection is usually one-way.

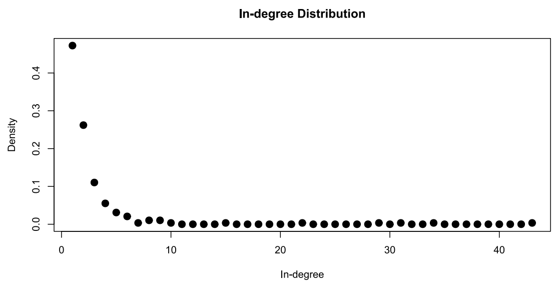


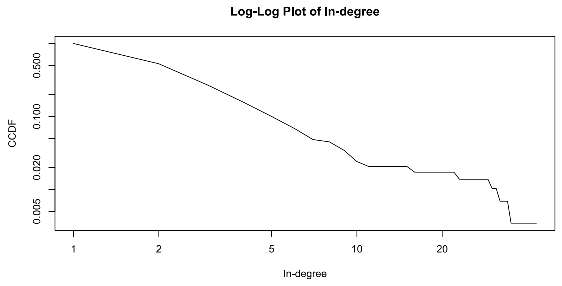
7. **(3 points) Present and interpret the in- and out-degree distribution based on your network as well as a log-log plot. Compute and interpret the estimate of the c slope (i.e., alpha value). Note that a p value (KS.p) less than 0.05 indicates the empirical data doesn't fit with the power-law distribution.**

Interpretation of c slope:

Here, we fit a power-law distribution to a vector containing samples from a power-law distribution. And in a power-law distribution, it is generally assumed that $P(X = x)$ is proportional to x^{α}, where x is a positive number and α is greater than 1. In practice, we may estimate the value of α with a given x_{min}.

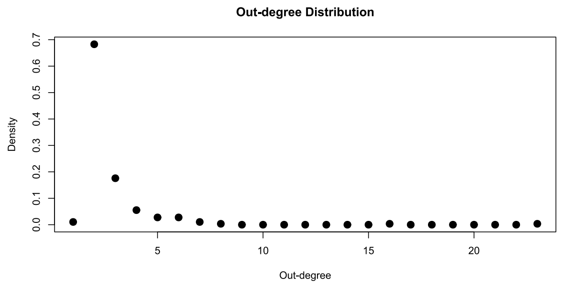
In-degree:

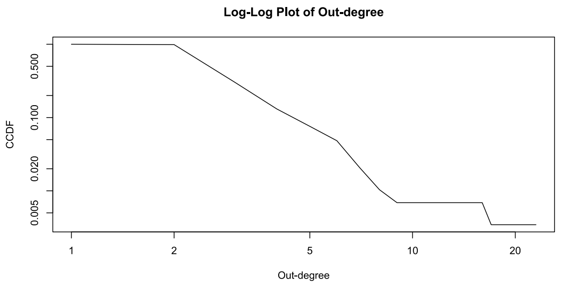




Continuous	α	x_{min}	logLik	KS.stat	KS.p
True	1.551833	0.01034483	11.81661	0.1462712	0.9955117

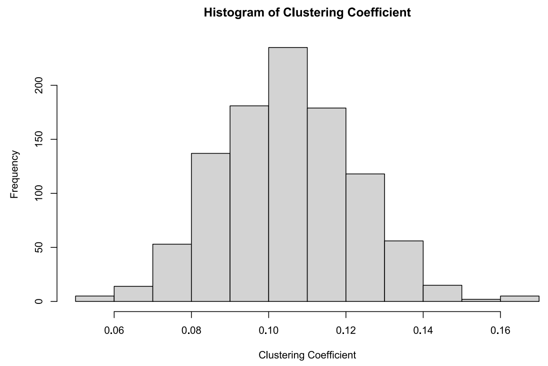
Out-degree:

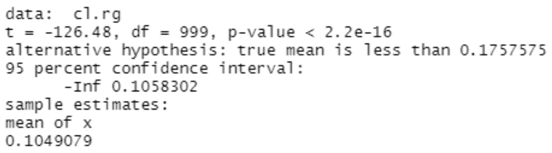




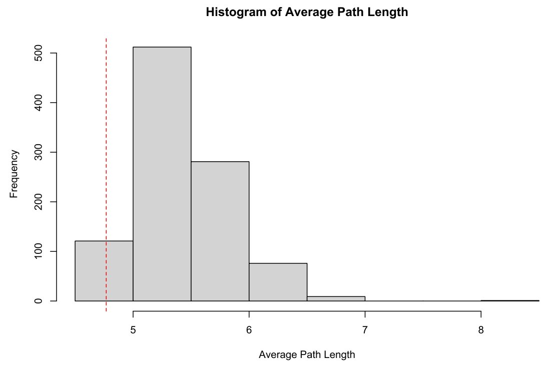
Continuous	α	x_{min}	logLik	KS.stat	KS.p
True	1.574659	0.01034483	11.33752	0.1891834	0.9636038

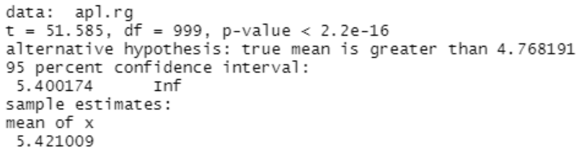
8. **(3 points) Present in a plot the observed and simulated values for each average path length and clustering coefficient based on the original network and 1,000 randomly shuffled networks.**





```
data:  cl.rg
t = -126.48, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is less than 0.1757575
95 percent confidence interval:
      -Inf 0.1058302
sample estimates:
mean of x
0.1049079
```





```
data:  apl.rg
t = 51.585, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 4.768191
95 percent confidence interval:
 5.400174      Inf
sample estimates:
mean of x
 5.421009
```

9. **(3 points) Based on these data would you conclude that the observed network demonstrates small world properties? If so, why? If not, why not?**

A network is said to be a small world only if its average clustering coefficient is **much greater** than that of a random graph constructed on the same set of nodes, and its average shortest path length is essentially **the same** as that of the random graph.

The observed network has average clustering coefficients of 0.1757575, which is larger than the mean of x, 0.1049079. Meanwhile, the observed average path length (with mean 4.768191) has almost the same distribution as the estimation (with mean 5.421009) Therefore, I believe that this observed network can properly demonstrates the small world properties

10. **(10 points) In two or three paragraphs, discuss your major findings of your network based on all the analyses you've done in this exercise and also your own additional analysis if necessary. Your answer here will be evaluated based on depth and comprehensiveness. Thus, you're encouraged to utilize extra information to answer this question. For instance, you can take a look at your original data (i.e., "twitterData," "youtubeData," or "redditData" if you work with the provided R code) in R. These data frames include additional user, text, and time information for your network. Similarly, if you need more insights from your network, feel free to run correlation and regression analysis based on your data collection.**

Evolution of Network: The figure2(2) below is the visualization result obtained from the second data collection on 2022/2/7. Compared with the data collected at the first time, which includes 290 nodes and 482 edges. The recently collected network data contains 476 nodes and 838 edges, and there exists an isolated sub-component in the network, which leaves 392 nodes and 698 edges after we extract the Giant Graph.

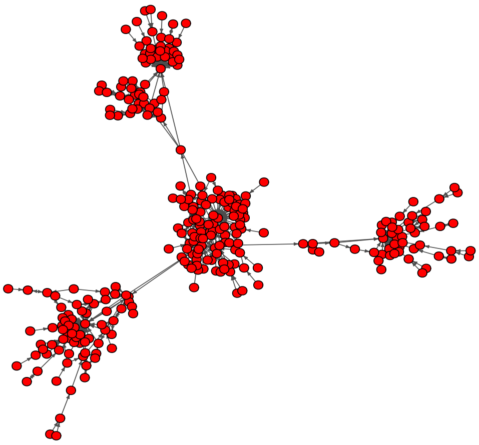


Figure 2(1) Data collected at the first time

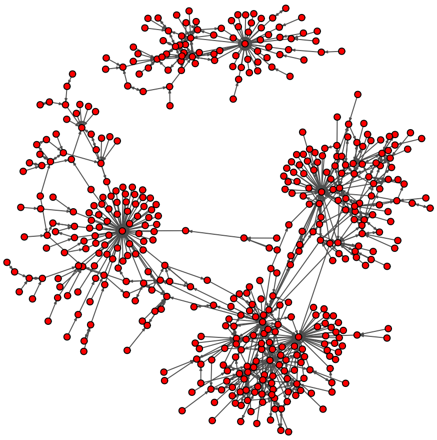


Figure 2(2) Data collected at the second time

After k-core decomposition visualization, the maximum k value of the network becomes 5, while the number of core components with large K value does not change significantly, only from 4 to 5. This shows that the social network has memory, and the later nodes will be affected by the existing nodes in the network. So it's easier to see pre-existing core component growth (i.e. k value increase) than to find a brand-new core.

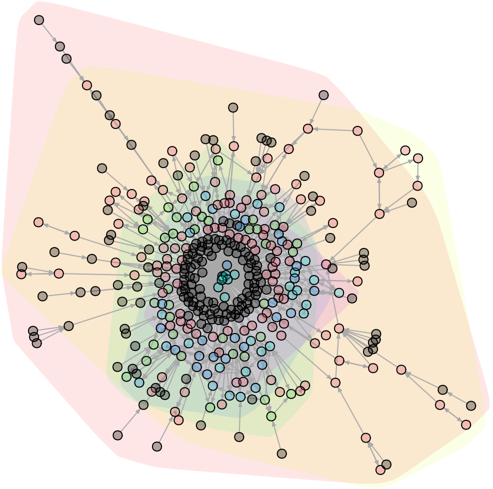


Figure 3(1) K-core decomposition

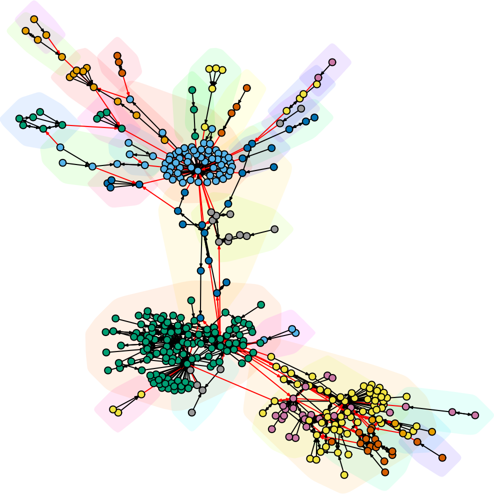


Figure 3(2) Cluster detection

Besides ,using the Walktrap algorithm, we find that the number of clusters in the network is reduced (from 32 to 30). This is because the newly added nodes connect some relatively isolated small clusters in the original network together, which reduces the number of small communities greatly. Therefore, while the total number of nodes increases, the number of clusters decreases.

Finally, it is not surprising that the observe Network does demonstrate some small world properties. After all, with the introduction of more nodes, the network structure will only become more stable.

K-core and Community: The number of communities found through the Walktrap algorithm is much larger than the number of k-core subgraphs, which means that many small communities are not included in these dense structures. Besides, we can only detect 9 communities when using Newman-Girvan algorithm, which indicates that edge-betweenness based algorithm has a low resolution compared with the other two community detection algorithms. As a result, it cannot separate out some small communities