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

Student: Yijian Li   Net ID: YLS9426

Part III Individual Network Properties:

1. 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.

Answer:

a): In-degree:

Index	Name	In-degree
35	BrainSnatch (139)	62
204	Resvrgam2 (1)	42
1	Perle1234 (251)	29
137	liverpooltarheels (287)	16
98	hacksoncode (128)	16

b): Out-degree:

Index	Name	Out-degree
98	hacksoncode (128)	19
35	BrainSnatch (139)	18
1	Perle1234 (251)	18
235	svmonkey (125)	12
171	NotCallingYouTruther (17)	12

c): Betweenness:

Index	Name	Betweenness
1	Perle1234 (251)	2.498938e+04
35	BrainSnatch (139)	1.867617e+04
15	armchaircommanderdad (26)	1.414513e+04
249	Titus_Favonius (166)	8.649167e+03
211	SAPERPXX (28)	6.994217e+03

d): In-closeness:

Index	Name	In-closeness
204	Resvrgam2 (1)	2.227668e-04
174	notwithagoat (89)	1.986887e-04

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

Student: Yijian Li   Net ID: YLS9426

220	Sirhc978 (50)	1.902226e-04
126	justonimmigrant (51)	1.808318e-04
137	liverpooltarheels (287)	7.747734e-05

e): Out-closeness:

Index	Name	Out-closeness
163	NapalmCheese (185)	2.661486e-05
134	LinguisticsIsAwesome (186)	2.643055e-05
16	asdfdasf98890_9897 (132)	2.634144e-05
40	BurtRogain (144)	2.629365e-05
211	Deep-Room6932 (153)	2.628466e-05

f): Eigenvector:

Index	Name	Eigenvector
36	Brandon_Beat_Trump (103)	1.000000e+00
68	do_you_even_ship_bro (244)	8.950891e-01
1	Perle1234 (251)	4.893438e-01
9	accountabilitycounts (241)	4.413985e-01
95	GreenAdvance (106)	2.629559e-01

g): Burt's network constraint:

Index	Name	Burt's network constraint
196	psunavy03 (45)	1.0306574
156	mossgiant95 (257)	1
87	freeloadingcat (258)	1
84	Footwarrior (267)	1
112	Impressive_Alarm_817 (123)	1

f): Hub score:

Index	Name	Hub
68	do_you_even_ship_bro (244)	1.000000e+00
9	accountabilitycounts (241)	4.449065e-01
95	GreenAdvance (106)	2.928312e-01
109	i_am_very_inteligent (240)	2.416275e-01
57	danmathew (243)	2.184539e-01

h): Authority score:

Index	Name	Authority score
36	Brandon_Beat_Trump (103)	1.000000e+00
1	Perle1234 (251)	5.927537e-01
6	A_Melee_Ensued (275)	7.841642e-02
191	Pliskin0331 (248)	4.709719e-02
17	AspiringArchmage (107)	3.482932e-02

2. 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...)?

Answer:

a): **In-degree:**

Degree centrality is the social networker's term for various permutations of the graph theoretic no- tion of vertex degree: for unvalued graphs, In-degree of a vertex, v, corresponds to the cardinality of the vertex set:

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

A **high in-degree** value such as BrainSnatch (139) mentioned above indicates that the node is more central, and this node is important in the whole graph with such the number of nodes it connects with.

b): **Out-degree:**

Out-degree corresponds to the cardinality of the vertex set

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

A **high out-degree** value such as hacksoncode (128) mentioned above shows this node is very active in contacting with other nodes and build a big network in all nodes.

c): **Betweenness:**

Betweenness takes one or more graphs (dat) and returns the betweenness centralities of posi- tions (selected by nodes) within the graphs indicated by g. Depending on the specified mode, betweenness on directed or undirected geodesics will be returned.

The shortest-path betweenness of a vertex, v, is given by

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

where g_{ivj} is the number of geodesics from i to k through j. Conceptually, **high-betweenness** vertices lie on a large number of non-redundant shortest paths between other vertices; they can thus be thought of as “bridges” or “boundary spanners.”

d): **In-closeness:**

Closeness takes one or more graphs (dat) and returns the closeness centralities of positions (se- lected by nodes) within the graphs.

In-closeness is the reciprocal of the average shortest length of a specific node connected **by** other nodes. In-closeness is from others and out-closeness is to others.

$$closeness(v) = \frac{1}{\sum_{i \neq v} d_{vi}}$$

The **high** value of in-closeness in my network could be expressed as how quick or how fast we could find **this node from other nodes** in the network as the information or message transmits thorough the best path.

e): **Out-closeness:**

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

$$closeness(v) = \frac{1}{\sum_{i \neq v} d_{vi}}$$

Out-closeness is the reciprocal of the average shortest length of a specific node connected **to** other nodes. In-closeness is from others and out-closeness is to others.

The **high** value of out-closeness in my network could be expressed as how quick or how fast we could find **other nodes from this node** in the network as the information or message transmits thorough the best path.

f): Eigenvector:

Eigenvector is a measure of the influence of a node in a network. The score here we applied is the relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores.

g): Burt's network constraint:

Burt's constraint is higher if ego has less, or mutually stronger related (i.e. more redundant) contacts. 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 in V[i], q != i,j)]^2, j in V[i], j != i).

for a graph of order (ie. number of vertices) *N*, where proportional tie strengths are defined as

p[i,j]=(a[i,j]+a[j,i]) / sum(a[i,k]+a[k,i], k in V[i], k != i)

a[i,j] are elements of *A* and the latter being the graph adjacency matrix.

For isolated vertices, constraint is undefined.

h): hub score:

The hub scores of the vertices are defined as the principal eigenvector of

$$R = A * A^T$$

where *A* is the adjacency matrix of the graph.

Hubs and authorities are a natural generalization of eigenvector centrality. There are two scores for each actor, a hub and an authority score. 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. These values are the same as the singular vectors derived from a single valued decomposition. There are scores associated with all the singular values but the one associated with the largest singular value is usually used as a centrality score.

h): Authority score:

The authority score of a vertex is therefore proportional to the sum of the hub scores of the vertices on the in-coming ties.

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

A **high** hub actor points to many good authories and a **high** authority actor receives from many good hubs.

3. **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**

Answer:

In graph theory and network analysis, indicators of **centrality** assign numbers or rankings to nodes within a graph corresponding to their network position. Applications include identifying the most influential person(s) in a social network, key infrastructure nodes in the Internet or urban networks, super-spreaders of disease, and brain networks.

Take NO.35 BrainSnatch (139) as an example:

	node_name	in_degree	out_degree	betweenness	incloseness	outcloseness	eigen	netconstraint	authority	hub
35	BrainSnatch (139)	62	18	1.867617e+04	6.276676e-05	2.578117e-05	7.650594e-03	0.02718726	7.824262e-03	1.459168e-03

As we can see, it has a very high out-degree and in-degree, as well as Betweenness, and hub value here.

So we can see the in and out degree is very essential in any network since they could transfer the information about how this node connects to others and whether this node plays an important role in the whole network like if it has a high out degree value, which kind represents many nodes in the network want to build a relationship with it or it has a high value for other nodes to find it.

Also, the betweenness is also top 5 in all nodes, which means that the number of times a node acts as a bridge along the shortest path between two other nodes. Therefore, we could know this node must be important in many shortest paths between other nodes, so it has a great value or potential for us to dig in. And this kind of trait could be observed from the high value of in or out degree.

What is more, we also find the authority and hub score is pretty high here. What we could know from the chart is that since BrainSnatch (139) has these 2 values in high, we conclude that it has good authority in the network with both high in-degree and out-degree, which let this node becomes authoritative and important center in network.

PART IV: Global Network Properties

1. **Briefly describe (a) what k-core is, (b) what insight this k-core decomposition method provides, and (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.)?**

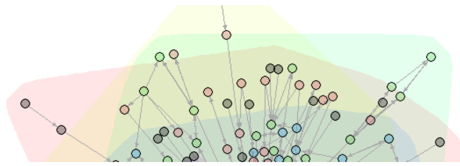
Answer:

a): k-core:

In general, a k-degenerate diagram is an undirected chart wherein each subgraph has a vertex of degree at generally k: that is, some vertex in the subgraph contacts k or less of the subgraph's edges. The decline of a diagram is the littlest worth of k for which it is k-degenerate. The decline of a diagram is a proportion of how inadequate it is and is inside a steady variable of other sparsity measures like the arboricity of a chart.

And we all know that:
A connected subgraph which has minimum degree greater than or equal to k, which means in the graph we draw, the different color represents different coreness and they are overlapped one by one with a increasing value of k-core value.

Like the colors below:



b): Insight:

So, from all these examples and out own graph, I know that we can find the largest core of the whole network and decompose it step by step to find all the subgraph it extends in lower layers.

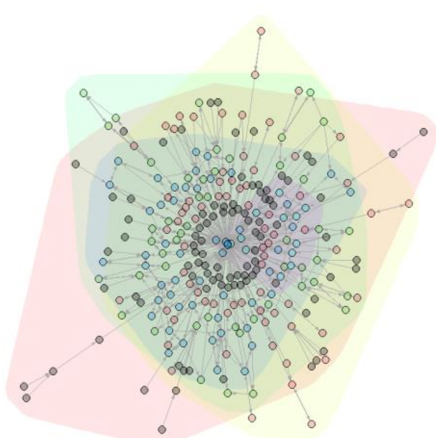
At the end, as we approach the lower layer, the nodes we meet will be more important as the more layers or cores they have interactions with.

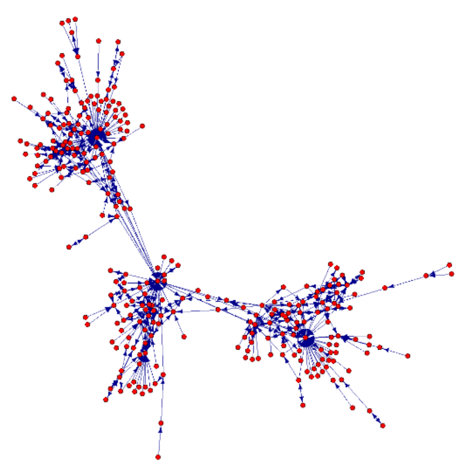
c):

The highest level of K I have in my network is 5. And there also several 4 and 3 in the network.

2. **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.**

Answer:





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

Student: Yijian Li   Net ID: YLS9426

In the left, I have the k-core decomposition one and also with the former graph we draw as comparison.

What we could observe form these two charts are that we see there are 4-5 centers in 3 different big clusters in the right graph, also we can see in the center of k-decomposition graph we also have 4-5 centers in the central point, which means they are the most important roles in the whole networks.

So, it is valid to conclude that the k-decomposition could enable us to enhance and verify the observation we did in lab 1 a with giant graph observation.

What is more, we can see there are 4-5 main colors in the left chart, so we could say the more central the nodes are, the more connections they have in the networks.

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

Answer:

IGRAPH clustering walktrap: groups: 26, mod: 0.71

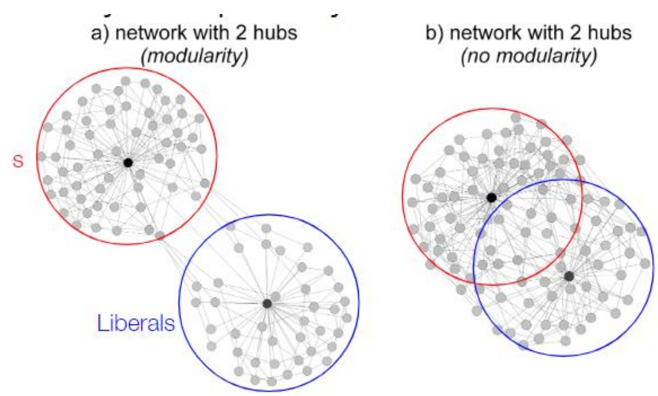
IGRAPH clustering edge betweenness: groups: 44, mod: 0.37

IGRAPH clustering info map: groups: 35, mod: 0.7

From the test results above, we can see the walktrap and info map are all good methods for implementing with nearly 70% modularity.

I choose walktrap method, since it has highest mod and lowest groups.

With these characteristics, we can know we could have more edges in communities than we expect by chance, like the chart below, the resulting graph will be more clear and easier for observation.



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

Answer:

There are **26** communities created and for a normal community might be the same reviews or comments related to the gun law problems in my network like safety problems or gun control problems.

5. **What is a modularity score? Interpret the modularity score of your results of community**

detection?

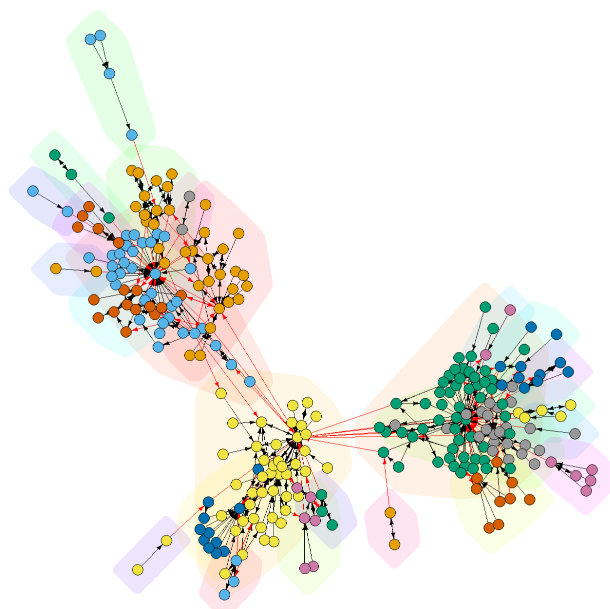
Answer:

With choosing cluster_walktrap method, we have modularity as 0.7119384.
This particular value represents a measure of the strength of division of a network into communities/subgraphs.

I think 0.7119384 is a pretty high value for general modularity, so I could say that the gun control is a very Gun control/law is a very iconic and polarizing topic, so the comments just polarized into their own communities or social groups.
With such a high value, we could also detect that when we faced with this kind of particular topic, it is not easy to change those people's mind in another community and they are very confident and do not easily waver in their beliefs and ideas.

6. **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?**

Answer:



We have the chart above to convey the information that there are mainly 3-4 groups (big communities) with different colors compositions. The communities are well separated and indeed there are some small overlaps in different communities.

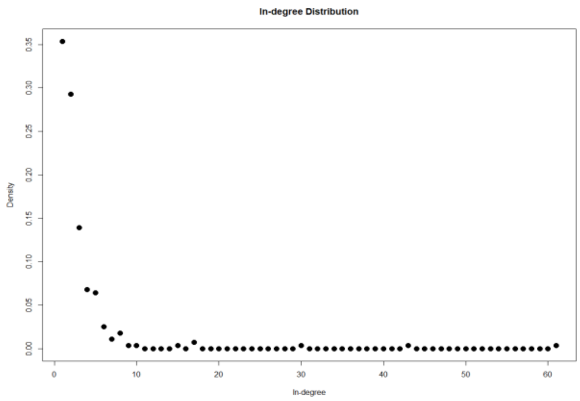
7. **Present and interpret the in- and out-degree distribution based on your network as well as a log-log plot. Compute and interpret the estimate of the c slope (i.e., alpha value). Note that a p value (KS.p) less than 0.05 indicates the empirical data doesn't fit with the power-law distribution.**

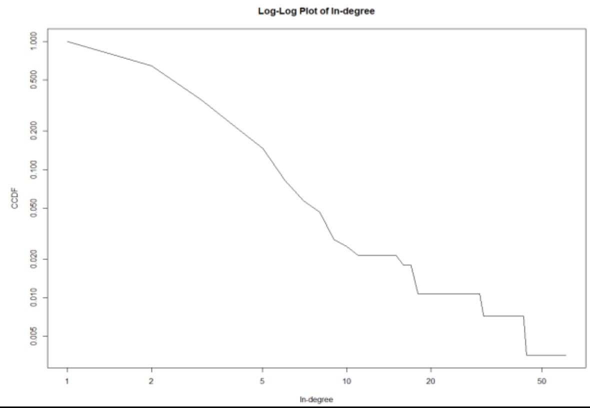
Answer:

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

Student: Yijian Li   Net ID: YLS9426

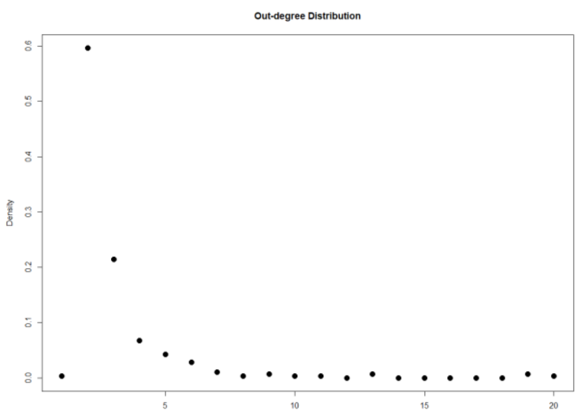
In-degree:

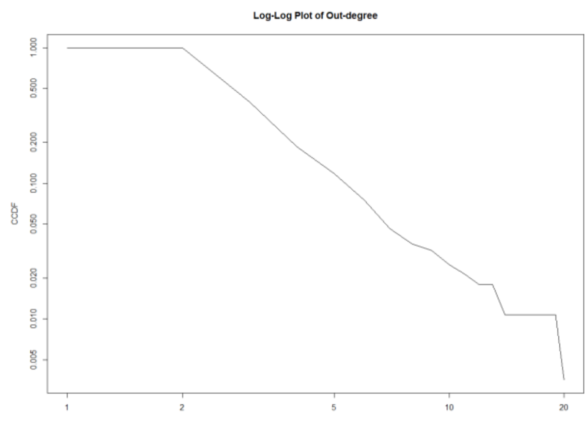




continuous	alpha	xmin	logLik	KS.stat	KS.p
TRUE	1.511102	0.007142857	11.82512	0.2302554	0.7265473

Out-Degree:





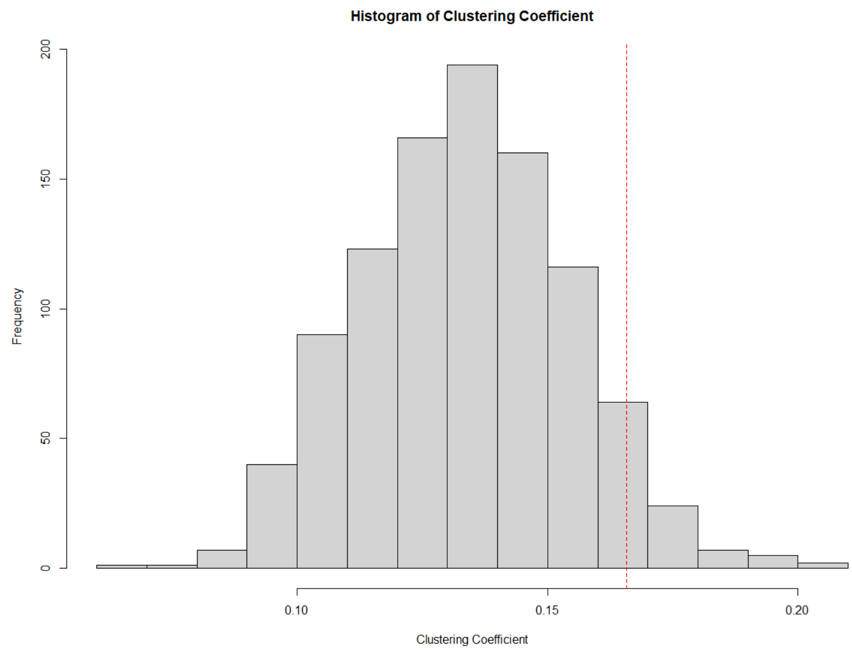
continuous	alpha	xmin	logLik	KS.stat	KS.p
TRUE	1.658874	0.02857143	5.200043	0.1967018	0.9903115

8. 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.

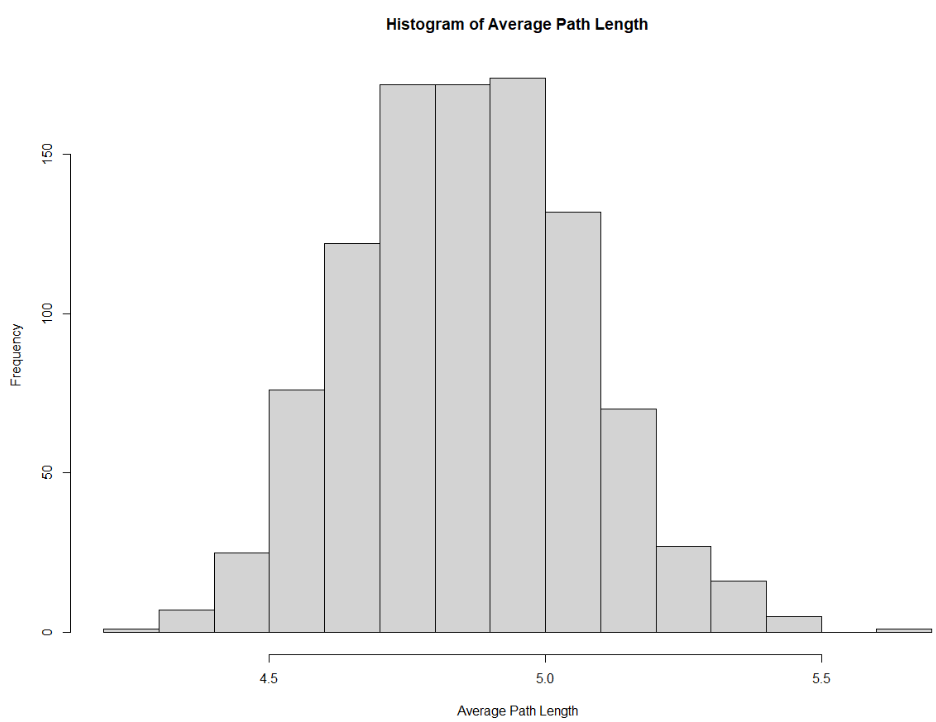
Answer:

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

Student: Yijian Li   Net ID: YLS9426



data: cl.rg
t = -49.239, df = 999, p-value = 1
alternative hypothesis: true mean is greater than 0.1657121
95 percent confidence interval:
0.1326145 Inf
sample estimates:
mean of x
0.1336854



data: apl.rg
t = -159.66, df = 999, p-value = 1
alternative hypothesis: true mean is greater than 5.909583
95 percent confidence interval:
4.848239 Inf
sample estimates:
mean of x
4.859072

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

Student: Yijian Li   Net ID: YLS9426

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

Answer:

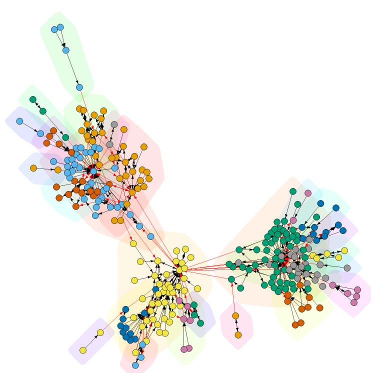
From my observations, if the global clustering coefficient of random networks greater than my networks, we could say that it is has small world properties.
However, as we can see the p-value above in Histogram of clustering coefficient is higher than the mean value in the chart, so we can say it **indeed demonstrates the small world properties**.
As the red line also show us about the global clustering coefficient of my network.

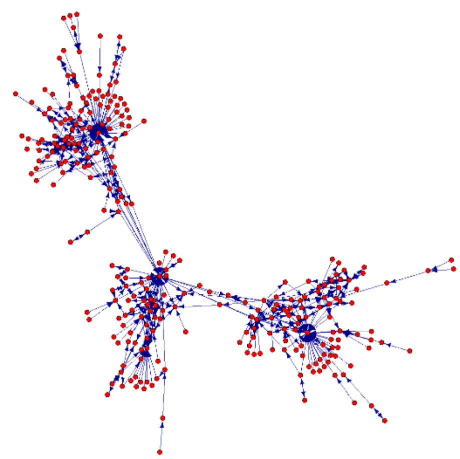
10. **In two or three paragraphs, discuss your major findings of your network based on all the analyses you've done in this exercise and also your own additional analysis if necessary. Your answer here will be evaluated based on depth and comprehensiveness. Thus, you're encouraged to utilize extra information 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.**

Answer:

Evaluation:

Basically, the data I choose are from reddit, a quite open platform for everyone to discuss, so the information I collected should be people's true ideas and suggestions.





Compared with the first time I draw the network picture, it is very unclear to see whether these nodes represent what kind of message and how to interpret them in a more general way like k-score or other parameters.

Also, as observing with idea of k-core, I understand why there are those layers or coreness and with deeper learning I figure out the relationship between number of k and the structure of the whole network.

So, it is valid to conclude that the k-decomposition could enable us to enhance and verify the observation we did in lab 1 a with giant graph observation, this is also one aspect of the improvement I achieve from lab1b.

At the same time, the idea of detecting algorithm also helps me how to see those graphs and know much clearer than before. In the three methods we learned, I also find out which kind of method could help me to obtain the best modularity to see how we can make the best communities division in our network.

Finally, as we made our assumptions about the indegree and outdegree valuations and those calculations in R, we made the conclusion that it could demonstrate the small world property.

Before attending this class, I have never thought about make such an interesting kind of research, all materials are collected by myself and make our own analysis about social network, which is great for me to start my learning.