COM 530500 **Network Science Final Project**

DUE: Thursday, January 20, 2022

No late homework will be accepted.

班級: 資應所

姓名: 鄭程哲

學號: 110065512

Problem 1. (40%) Consider the **ego-Facebook** [1] dataset. A node in this dataset represents a user on Facebook, and an edge between two nodes represents the relationship between two users.

(a) (10%) List some statistical information of this dataset, such as the number of nodes, number of edges, average clustering coefficient, diameter, average degree, maximum degree, etc.

(b) (10%) Visualize the dataset by plotting it.

(c) (10%) Plot the degree distribution with log-log scale.

(d) (10%) List the top 10 nodes ranked by the following centrality measures.

- Degree centrality
- Katz centrality
- Eigenvector centrality
- Betweenness centrality
- Closeness centrality

Solution:

(a) Statistical information:

- The number of nodes: 2851
- The number of edges: 62318
- Average clustering coefficient: 0.591376
- Diameter: 14
- Average degree: 43.7166
- Maximum degree: 769
- Minimum degree: 1
- Density: 0.01534

(b) Visualization is as figure 1 show.

(c) Degree distribution is as figure 2 show.

(d)
- Degree centrality: 1165 91 0 1862 1681 1606 288 1341 1575 1298

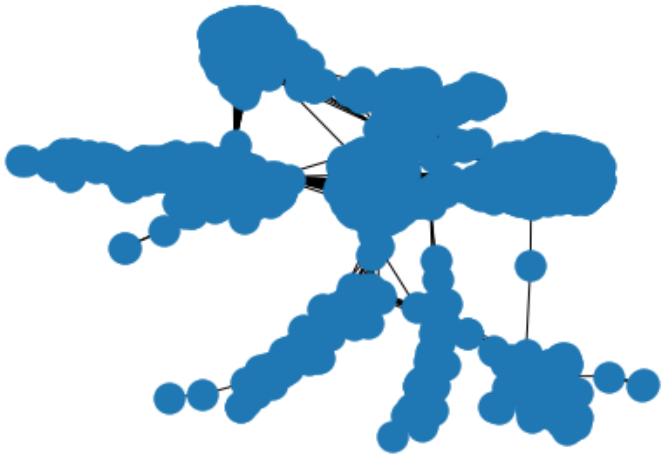


Figure 1: Visualization

- Katz centrality: 1165 91 1862 1681 0 1606 1341 1575 1298 1493
- Eigenvector centrality: 1606 1551 1575 1790 1431 1561 1493 1741 1581 1349
- Betweenness centrality: 91 1165 0 1110 611 1056 771 892 1181 351
- Closeness centrality: 91 1165 1156 1086 933 1110 1181 1056 147 49

◇

Problem 2. (60%+bonus 10%) In this problem, we want to investigate the disease propagation by the independent cascade (IC) model in **ego-Facebook** [1] dataset. Assume the propagation probability is ϕ, and the set of seeds nodes S are randomly selected. Collect the set of infected nodes within the distance D of the seed nodes, and calculate the prevalence rate r_1 (which is defined by the ratio of the number of infected nodes to the total number of nodes). Set $\phi = 0.1, |S| = 5$, and D the diameter of the graph.

(a) (40%) Simulate the disease propagation by IC model after removing the top 0%, 10%, 20%, ..., 50% of nodes from the following centrality measures respectively, and calculate the corresponding prevalence rate r_1. Please plot the curves of r_1 vs. the percentage of nodes removed. (*Note: Please run the simulation 100 times and average the results.*)

- Degree centrality
- Katz centrality
- Eigenvector centrality
- Betweenness centrality
- Closeness centrality

(b) (bonus 10%) Could you find a centrality measure that achieves a better performance?

(c) (20%) Write a report to compare and discuss the results of different centrality measures.

Solution:

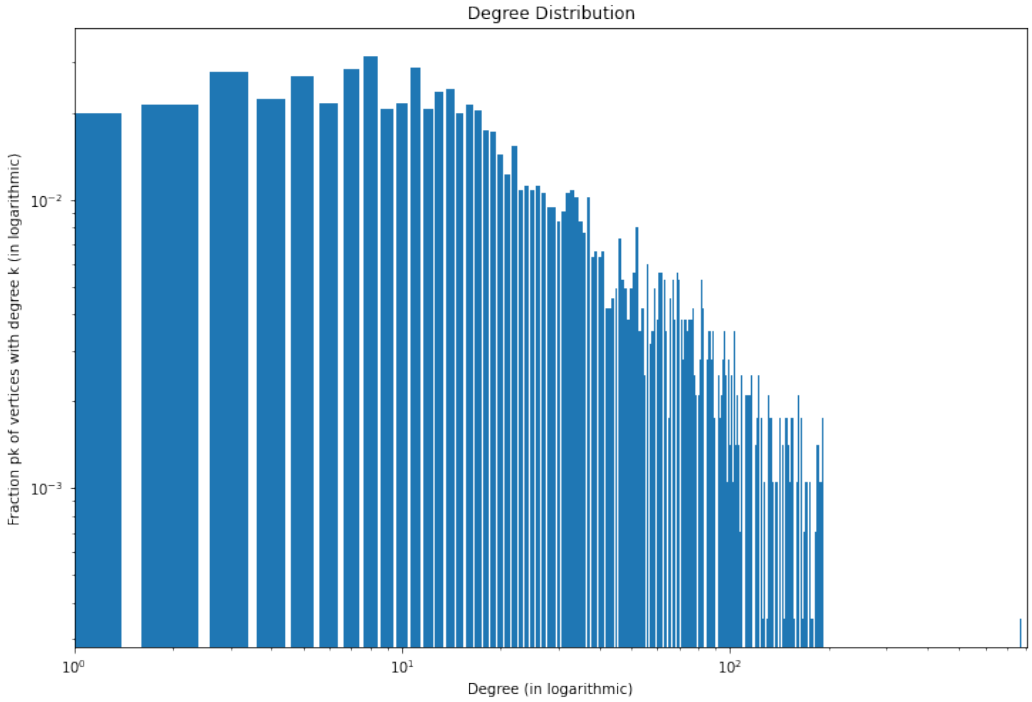


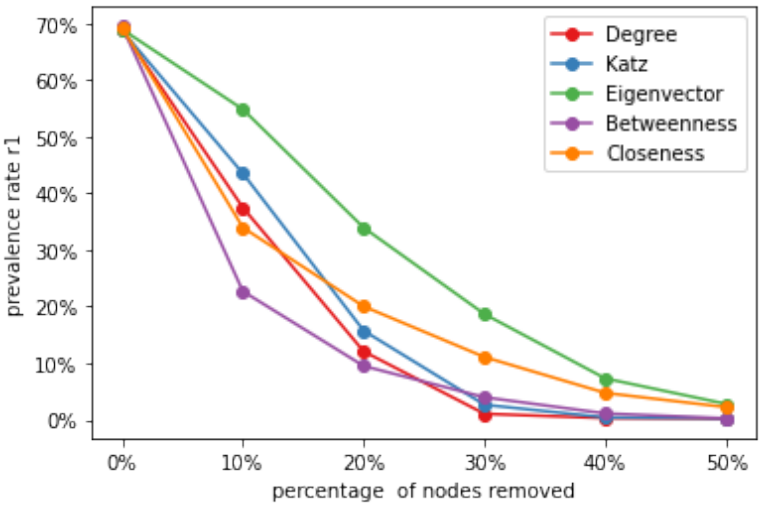
Figure 2: Degree distribution

(a) The experiment results are shown as the following table and figure 3. The bold text in the table means the lowest $r1$ comparing to the other measures. In figure 3(b), I additionally plot the box information corresponding to different measures.

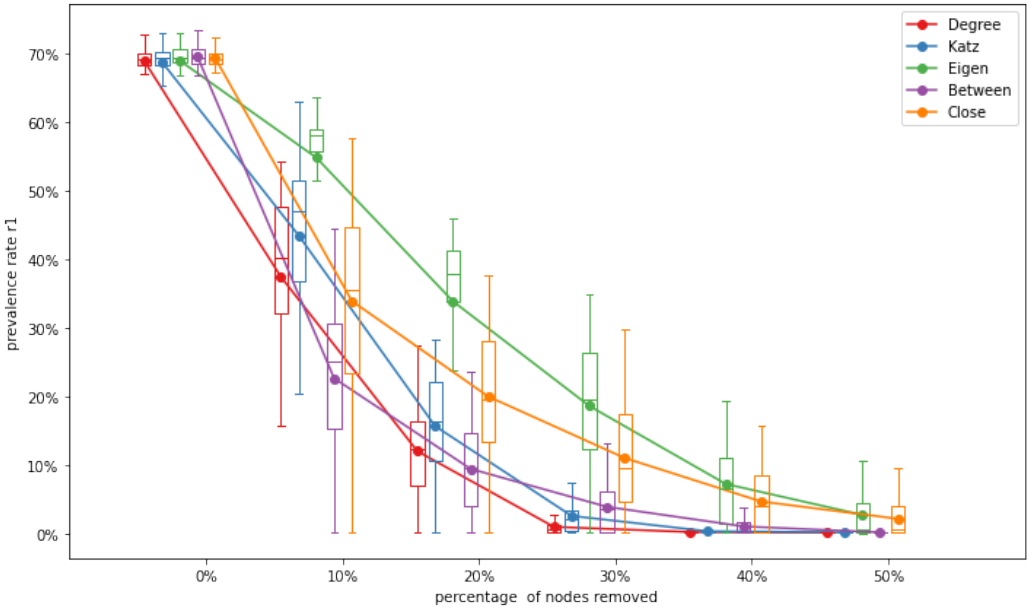
Measure	0%	10%	20%	30%	40%	50%
Degree	0.689032	0.374749	0.120698	**0.010463**	**0.002925**	**0.002066**
Katz	**0.687464**	0.435170	0.157306	0.026429	0.004037	0.002413
Eigenvector	0.689747	0.548369	0.339309	0.186615	0.072929	0.027938
Betweenness	0.695184	**0.226962**	**0.094949**	0.039646	0.011122	0.002434
Closeness	0.693094	0.339548	0.200284	0.110716	0.047341	0.022210

(b) I tried other different centrality measures including PageRank, VoteRank, and Harmonic centrality. The experiment results are shown as figure 4. As we can see from the table below, PageRank breaks two best records, which are 20% and 50%, respectively. I think PageRank is a quite good centrality measure. Further discussion would be written in the part (c). In contrast, VoteRank and harmonic centrality measures didn't have a better performance.

Measure	0%	10%	20%	30%	40%	50%
Best from (a)	**0.687464**	**0.226962**	0.094949	**0.010463**	**0.002925**	0.002066
PageRank	0.689937	0.276692	**0.084177**	0.015167	0.004504	**0.001884**
VoteRank	0.688888	0.280277	0.086461	0.040621	0.008362	0.009404
Harmonic	0.694942	0.333560	0.176552	0.092083	0.033385	0.018622

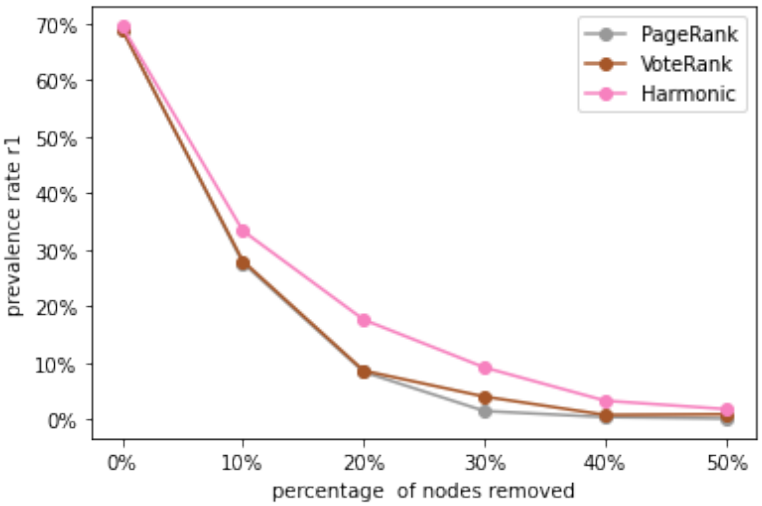


(a) Line chart

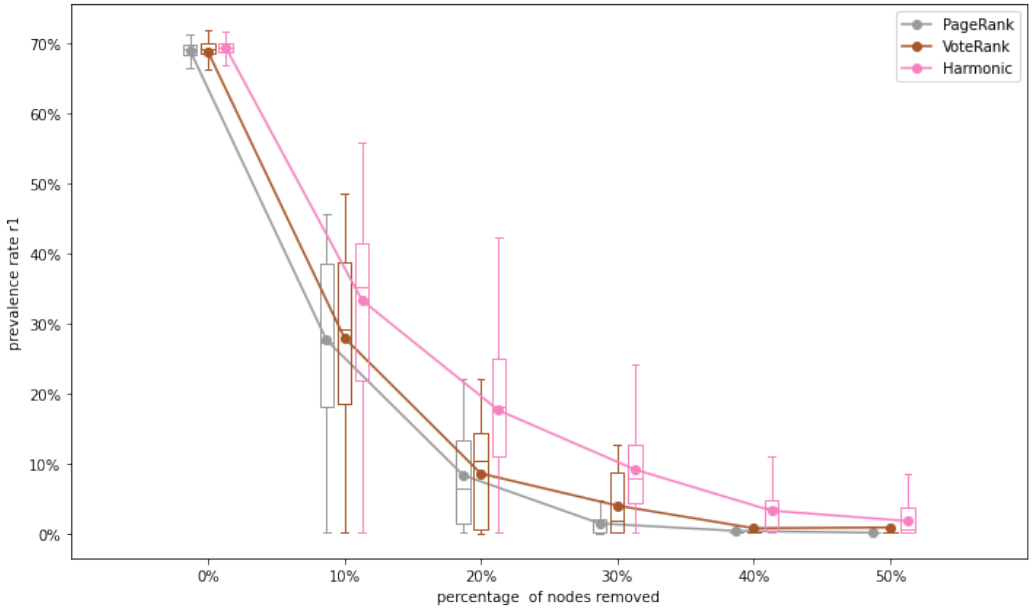


(b) Line chart with box

Figure 3: $r1$ curve corresponding to different measures and percentage



(a) Line chart



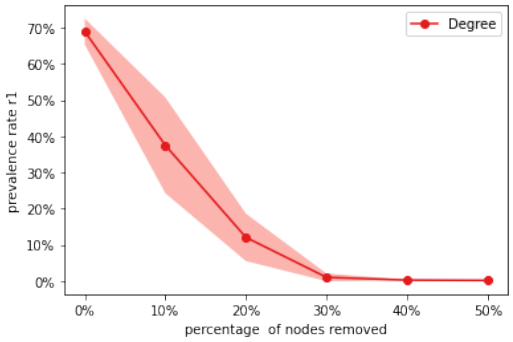
(b) Line chart with box

Figure 4: $r1$ curve corresponding to other measures

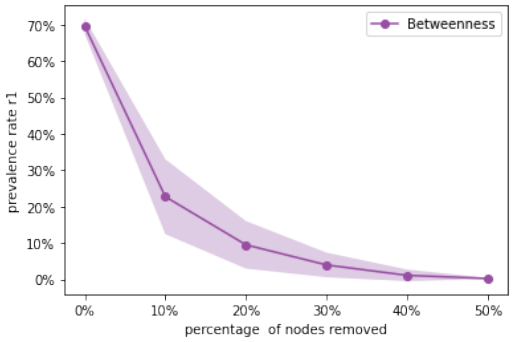
(c) Removing 0% vertices has no difference for different measures, so we can ignore this column. When the percentage of node removal is low (10%, 20%), $r1$ drops a lot with betweenness, PageRank, and VoteRank measures. That is, applying this concept into real world, we can use these 3 measures to reduce $r1$ efficiently when the shortage of vaccine occurs.

For another thing, degree centrality achieves 2 best $r1$, betweenness centrality achieves 1 best $r1$, and PageRank achieves 2 best $r1$ in total. In figure 5, I plot more details about these 3 good measures. The dot in the plot means the average value of $r1$, while the shaded area represents ± 1 standard deviation of mean. As we can see from the figure, the higher percentage of nodes removal, the performance converge more tightly. In other words, the variance of performance is higher if the shaded area is bigger. The performance with high variance would be not accurate. Although both degree centrality and PageRank has 2 best $r1$ records, PageRank achieves the best performance in 50% nodes removal.

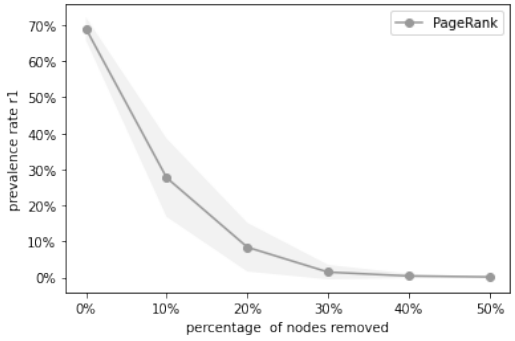
Combining above two points, I, therefore, think PageRank is the best centarlity measures among all in this task.



(a) Degree centrality



(b) Betweenness centrality



(c) PageRank

Figure 5: $r1$ curve corresponding to a specific measure

◇

References

[1] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems*, 2012, pp. 539–547.