Report

Xiaohui Li, Luhuan Wu

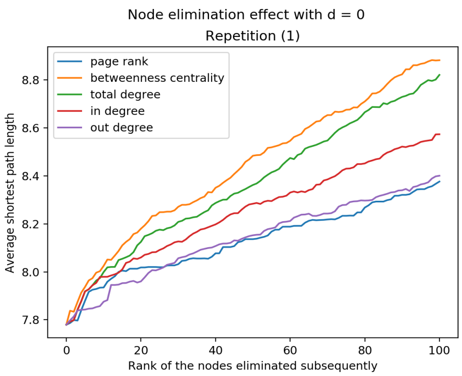
1 Effect on average shortest path

In this section, with the effective tool directed configuration model, we focus on the ranking algorithm's measure of graph's connectivity efficiency. Concretely, for a graph G, we choose the average shortest path length l_G as the indicator of connectivity efficiency. And we marginally eliminate the individual node based on different ranking algorithm and see the cumulative effect upon the average short path. Note that each time we eliminate a node, we use the remnant graph's strongly connected component to calculate the new average shortest path length.

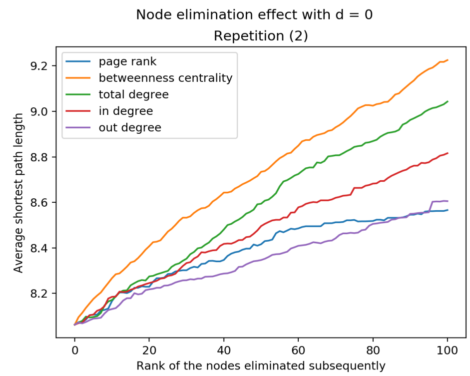
1.1 Node elimination effect

We study the effect of dependence between in-degree and out-degree to the ranking measure performance in average shortest path length. We vary the parameter d and fix the other parameters to control for the random graph features, and we then use the path length against deleted node number plot to analyze the effect. Specifically, we set the power index $\alpha = 2.1$, $\beta = 3$, the expected degree mean E = 3, the node size of graph n = 5000, and we consecutively eliminate 100 nodes.

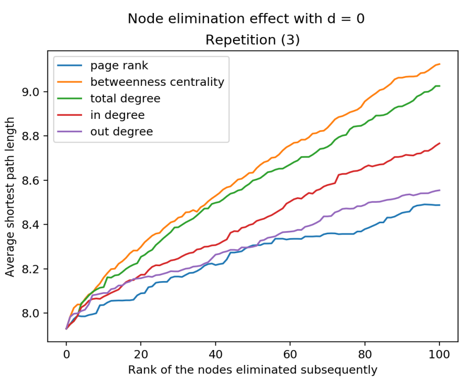
Due to the randomness from random graph, there is noise in the simulated graph's shortest average path length. Therefore we repeatedly conduct the simulation and take the average to smooth out the noise, thus deriving the mean node elimination effect on path length. In the following graphs in next page, we could clearly see the noise of shortest average path length.



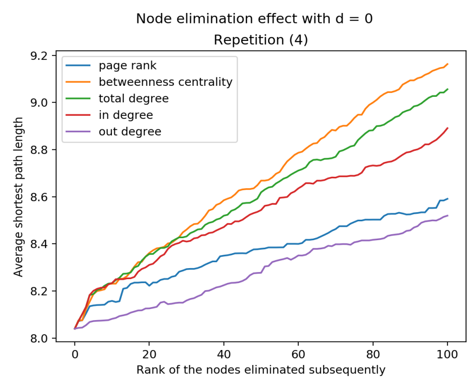
(a) Repetition(1)



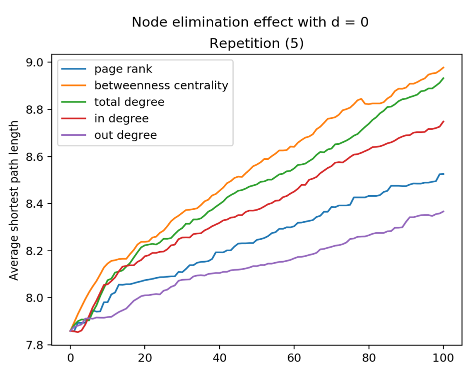
(b) Repetition(2)



(c) Repetition(3)



(d) Repetition(4)



(e) Repetition(5)

Figure 1.1: Repetitive model simulation of d = 1 case

We take the sample mean of these five samples and get the average effect on shortest path length. It could be seen that the average smooth out the noise, the figure is shown as below.

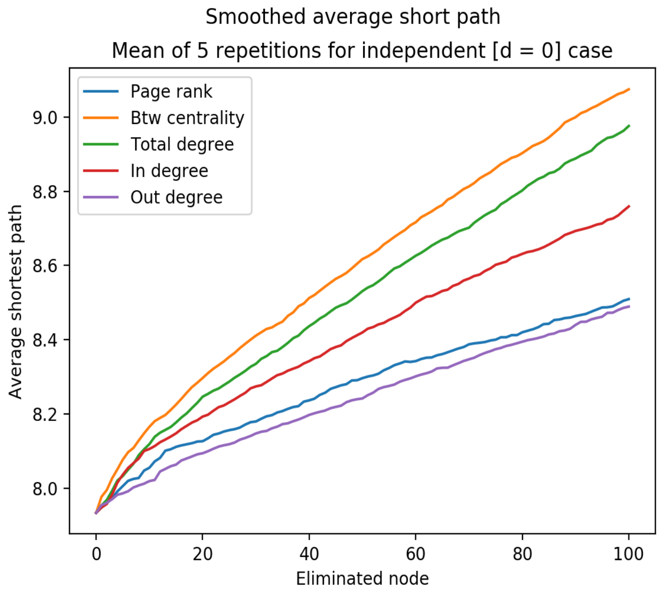


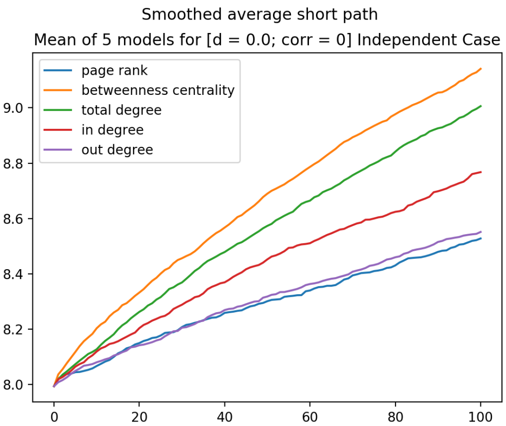
Figure 1.2: Average effect on average shortest path length

Also we could find that in the independent case, betweenness centrality has the most effect on average shortest path length, with total degree slightly less than it, this denotes that the graph's average shortest path length is most sensitive to the betweenness centrality, corresponding to the centrality feature betweenness centrality is designed to measure. However, PageRank is far worse than those two and in-degree, it seems that PageRank fails to capture the average shortest path length characteristics.

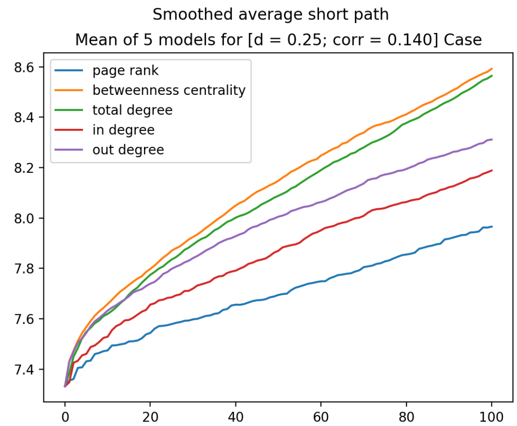
1.2 Dependence influence on the effect

Further we check the effect in different dependence level, we assign the d with 0, 0.25, 0.5, 0.75 and 1 five levels, and check the results. To smooth out the noise, each dependence level case is also simulated by five replications and then taking the average.

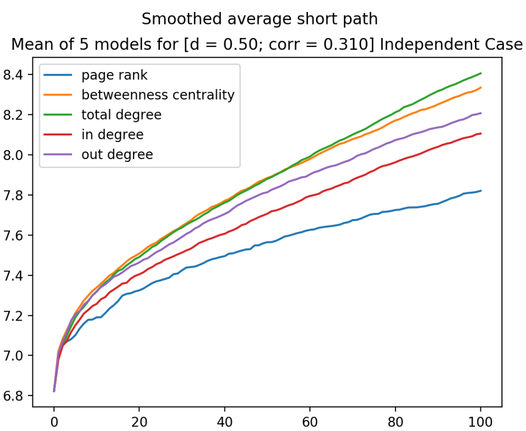
The results shows that as the dependence grows, the total degree gradually outperforms the betweenness centrality. And the PageRank is the worst regardless of the degree dependence. This suggests that PageRank fails to capture the graph communication efficiency (average shortest path length), while the betweenness centrality and the total degree both measure well. In the case where in-degree and out-degree are highly correlated, the simple total degree ranking is even better. The respective graphs are in the next page.



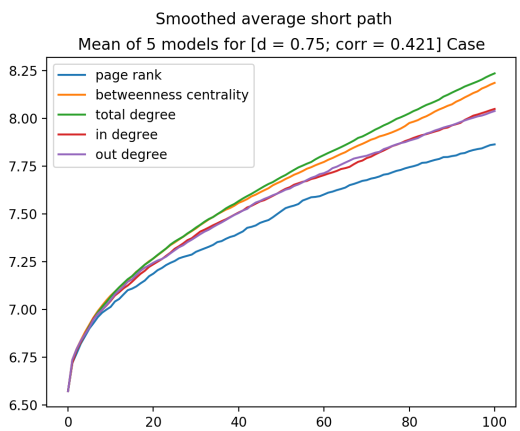
(a) d = 0 independent case



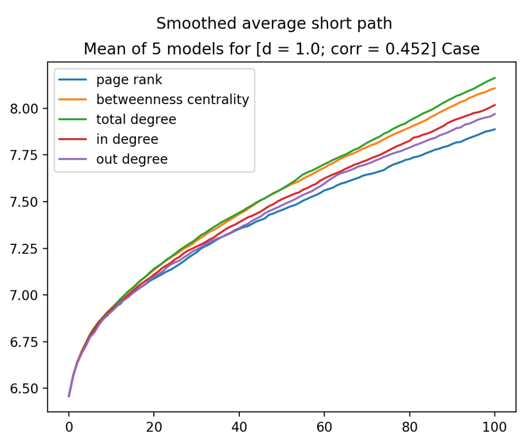
(b) d = 0.25 case



(c) d = 0.5 case



(d) d = 0.75 case



(e) d = 1.0 case

Figure 1.3: Effect on average shortest path length in different dependence

2 Real world data

We apply our analysis into real world data, and we get corresponding findings.

In consideration of the directed configuration graph model feature, we select the Wikipedia vote network from Stanford Large Network Dataset Collection, which is a directed graph with in-degree and out-degree following the power law distribution.

The real data statistics is shown in the following table.

Table 1: Wikipedia vote network statistics

Feature name	Value
Nodes size	7115
Edges	103689
Expected degree	14.57
In-out degree correlation	0.317
Average clustering coefficient	0.1409
Number of triangles	608389
Diameter(longest shortest path)	7

The node elimination effect on real graph's average shortest path length is shown below.

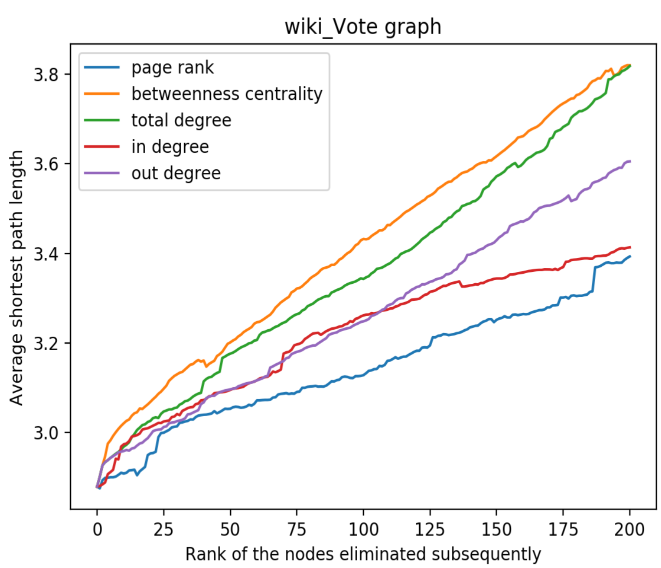


Figure 2.1: Real graph average effect on average shortest path length

The graph shows a consistent result with previous simulation results. PageRank is the worst in measuring the significance of average shortest path length. And both the betweenness centrality and total degree perform well with betweenness centrality slightly better.

We fit the real data with previous model. Concretely, we fit the power law index from empirical log-log distribution and then apply linear regression in the power law shape component to get the slope, thus α and β. We also get the fitting expected degree E, graph size n from real data feature, and dependence factor d from the inverse calculation of real data degree correlation. The parameters are: $\alpha = 3.5$, $\beta = 2.5$, E $= 14.57$, d $= 0.339$ and n $= 7115$.

Yet there exists difference between fitting model and real graph. Especially the clustering coefficient, the real data has average clustering coefficients of 0.1409, while the simulation graph is just 0.004693. The real graph is actually a small-world whereas the directed configuration graph fails to capture.

The simulation model graph is shown below, as before, we take the average from repeated model simulation average shortest path length to reduce the noise.

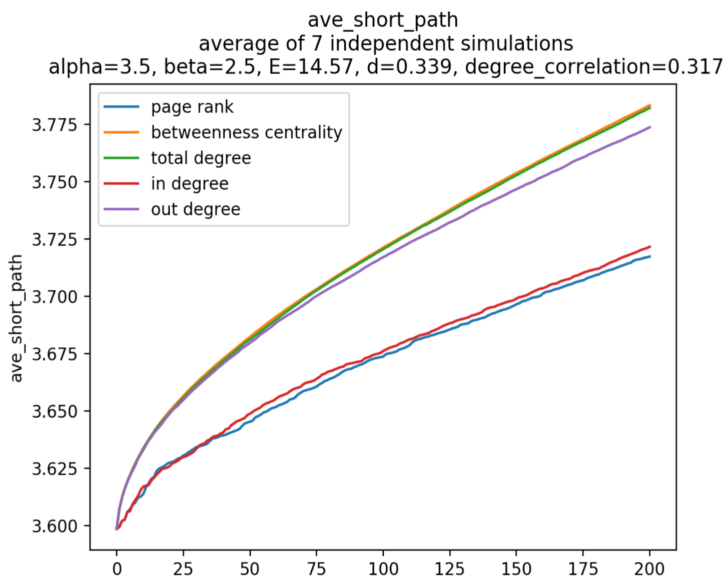


Figure 2.2: Fitting model average effect on average shortest path length

The result is generally similar with the real data with same trend and ranking algorithm performance. Nevertheless, there is some difference, the total degree is the same as betweenness centrality while in real graph betweenness is better, besides, the real graph has larger path length range (2.8 to 3.8), while the simulated mean only increases from 3.6 to 3.8. The potential explanation may be because the simulated graph is the average case, reflecting the expected effect, yet the real graph is a sample realization, it has more noise.