Analysis of Relationship between PageRank and Betweenness Centrality, and Effects on Average Shortest Path Length

Luhuan Wu, Xiaohui Li

May – Sept. 2017

1 Abstract

In this study, we focus on two aspects: how PageRank (PR) and Betweenness Centrality (BC) are correlated, and how measures including PR, BC and degree measures (total degree, in-degree, out-degree) evaluate the graph's connectivity efficiency, of which we choose Average Shortest Path Length(ASPL) as an indicator. Specifically, we investigate the effects of in-degree–out-degree correlation on the study objects above. Using analytical treatments and numerical simulation of Directed Configuration Models, we show that for the first study, high dependence of in- and out- degree sequence enhances the correlation between PR and BC; as for the second study, BC and total degree mostly affect ASPL, while PR is the worst indicator, with in-degree and out-degree having medium effect; furthermore, as degree dependence increases, BC's performance on ASPL declines while total degree still holds a dominant place, and the difference among performance of the five measures narrows. This suggests that total degree could be a reliable measure in evaluation of ASPL. Lastly, we extend our study to a real-world directed graph–Wiki-vote network, and show that the results are consistent.

2 Description

Our simulation studies are divided into two parts:

1. Relationship between PageRank (PR) and Betweenness centrality (BC).

2. Comparison among Effect of Measures (PageRank, Betweenness centrality, Total degree, In-degree and Out-degree) on Average Shortest Path Length (ASPL)

In Part 1, we want to study to what extent PR and BC are evaluating the same 'importance / centrality' of a node. Therefore, we alter the degree correlation to see how does the relationship between PageRank and Betweenness centrality change accordingly.

In Part 2, we want to find out how the ASPL value would change, if we remove the node of highest Page rank, and then the second highest, and so on; and how it behaves if we do this according to the ranking of Betweenness centrality, Total degree, In-degree, or Out-degree.

The models in this simulation projects are mainly implemented by Directed Configuration Model (DCM). In addition to that, we analyze the real-world graph *Wiki-Vote network*, and simulate this network using DCM and then apply similar treatments.

A thorough report on simulation and analysis is given below. Specifically, links of output data are attached to corresponding experiment, and more detailed information of data and codes availability is in Appendix.

3 Model specifications

3.1 Directed Configuration Model

3.1.1 DCM and Algorithm

The detailed description of the DCM is given in [1]. We implement both *Repeated directed configuration model* and *Erased directed configuration model* described in [1]. However, after testing, we find that erased algorithm is much faster than repeated algorithm, so in following simulations we adopt the erased algorithm to generate DCM.

3.1.2 Programming and Package

We use Python as our programming language due to its supportive community. The full Python code of this project, from building the graph, testing and analyzing are in the project's GitHub page [2].
The package that helps us most to generate the model is *networkx 1.11* [3] and the main methods implemented are:

1. directed_configuration_model [source code]

   ```
   directed_configuration_model(in_degree_sequence,
       out_degree_sequence, create_using=None, seed=None)
   ```

   Return a directed_random graph with the given degree sequences.

   The configuration model generates a random directed pseudograph (graph with parallel edges and self loops) by randomly assigning edges to match the given degree sequences.

   To remove parallel edges:

   ```
   >>> D=nx.DiGraph(D)
   ```

   To remove self loops:

   ```
   >>> D.remove_edges_from(D.selfloop_edges())
   ```

2. pagerank [source code]

   ```
   pagerank(G, alpha=0.85, personalization=None, max_iter=100, tol=1e
       -06, nstart=None, weight='weight', dangling=None)[
   ```

   Return the PageRank of the nodes in the graph.

   PageRank computes a ranking of the nodes in the graph G based on the structure of the incoming links. It was originally designed as an algorithm to rank web pages.

   Notes:

   The eigenvector calculation is done by the power iteration method and has no guarantee of convergence. The iteration
   [5] [6]

3. betweenness_centrality [source code]

   ```
   betweenness_centrality(G, k=None, normalized=True, weight=None,
       endpoints=False, seed=None)
   ```

Compute the shortest-path betweenness centrality for nodes.

Betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

where V is the set of nodes, $\sigma(s,t)$ is the number of shortest $(s,t)-$paths, and $\sigma(s,t|v)$ is the number of those paths passing through some node v other than s, t. If $s = ts = t, \sigma(s,t) = 1$, and if $v \in s, t, \sigma(s,t|v) = 0]$.

Notes:

The algorithm is from Ulrik Brandes [7]. See [**?**] for the original first published version and [8] for details on algorithms for variations and related metrics.

For approximate betweenness calculations set k=#samples to use k nodes ('pivots') to estimate the betweenness values. For an estimate of the number of pivots needed see [10].

For weighted graphs the edge weights must be greater than zero. Zero edge weights can produce an infinite number of equal length paths between pairs of nodes.

4. average_shortest_path_length [source code]

```
average_shortest_path_length(G, weight=None)
```

Return the average shortest path length.

The average shortest path length is

$$a = \sum_{s,t \in V} \frac{d(s,t)}{n(n1)}$$

where V is the set of nodes in G, $d(s,t)$ is the shortest path from s to t, and n is the number of nodes in G.

3.1.3 Theoretical Model

$$P(D^+ = d^+, D^- = d^-) = P(Poisson(W^+) = d^+, Poisson(W^-) = d^-)$$

where (W^+, W^-) are jointly regularly varying.

The distribution of (W^+, W^-) is given as follows:

Marginal distribution: W^+ and W^- both have power-law distribution, where

$$\lim_{x \to \infty} P(W^+ < x) = 1 - (\frac{x}{b})^{-\alpha}$$

$$\lim_{x \to \infty} P(W^- < x) = 1 - (\frac{x}{c})^{-\beta}$$

Joint distribution:

$$W^+ = ad(W^-)^S + a(1-d)(W^-)^s$$

where $\hat{W}^-$ is an independent copy of W^-.

- Dependencies of parameters

$$
\begin{aligned}
s &= \frac{\beta}{\alpha} \\
a\frac{\beta c^s}{\beta - s} &= \frac{\beta c}{\beta - 1} \\
\Rightarrow a &= 1 \text{ when } s = 1 \\
a &= \frac{\beta - s}{a^{s-1}(\beta - 1)}, \text{ otherwise}
\end{aligned}
$$

  Proof. in Appendix

  The degree of freedom is 4. In the following simulation, we choose the values of α, β, d and expected degree to fix the model.

- Further calculations of expected degree and degree correlation

  - Expected Degree

$$
\begin{aligned}
p_{w^-}(x) = PDF(W^-) &= \beta c^\beta x^{-\beta-1} \\
\Rightarrow E[(W^-)^s] &= \int_c^\infty \beta c^\beta x^{-\beta-1} x^s dx \\
&= \frac{\beta c^s}{\beta - s} \\
E[(W^-)^{2s}] &= \frac{\beta c^{2s}}{\beta - 2s} \\
E[(W^-)^2] &= \frac{\beta c^2}{\beta - 2} \\
E[W^-] &= \frac{\beta c}{\beta - 1}
\end{aligned}
$$

    Since $s = \frac{\beta}{\alpha}$, by setting $\alpha, \beta > 2$, the denominator would be non-zero.

  - Degree Correlation

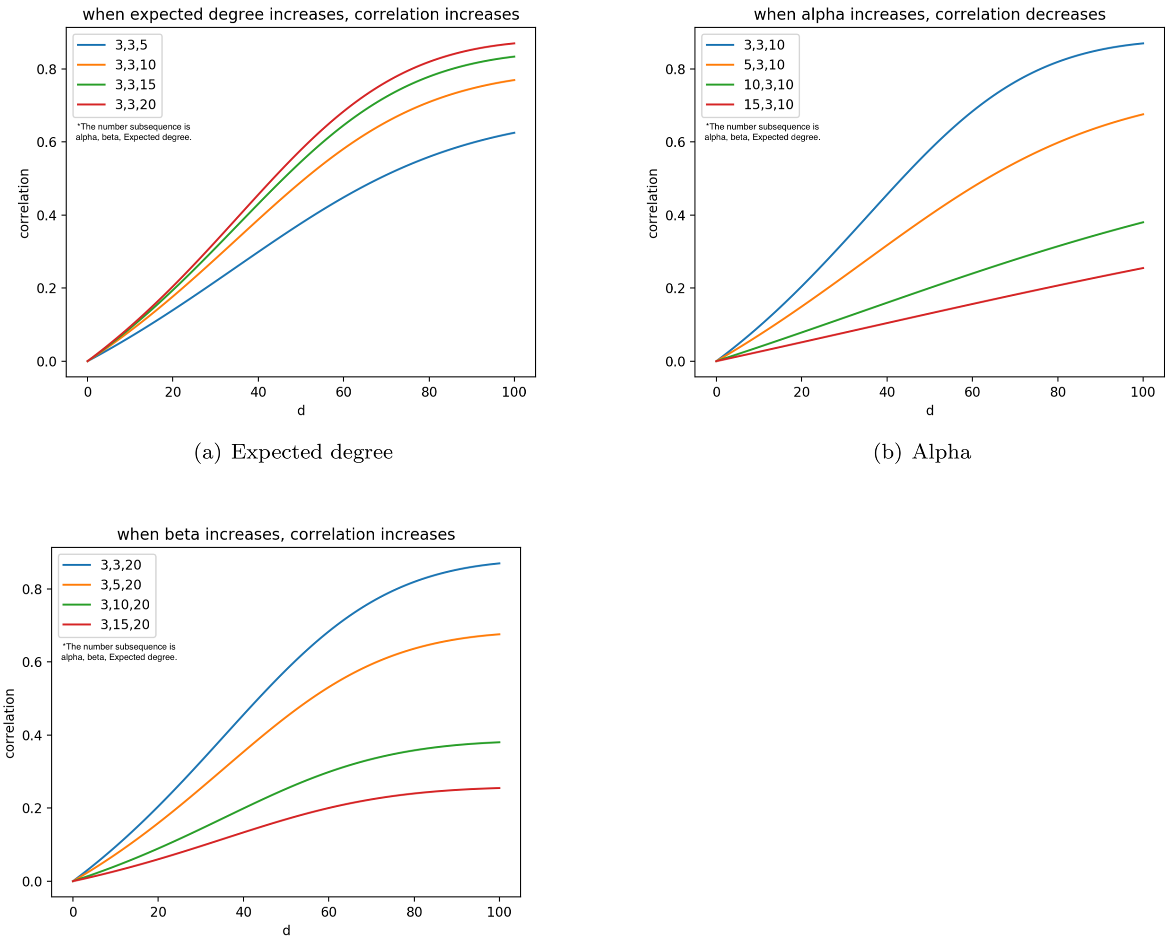
$$
\begin{aligned}
Var(D^+) = E[D^+] - (E[D^+])^2 = E[W^+ + (W^+)^2] - (E[W^+])^2 \\
E[W^+] = aE[(W^-)^s] \\
E[(W^+)^2] = a^2(d^2 + (1-d)^2)E([W^-)^{2s}] + 2a^2 d(1-d)(E[(W^-)^s])^2
\end{aligned}
$$

    Similarly,

$$
Var(D^-) = E[D^-] - (E[D^-])^2 = E[W^- + (W^-)^2] - (E[W^-])^2
$$

    Thus we can obtain the covariance of D^+ and D^-:

$$
\begin{aligned}
Cov(D^+, D^-) &= E[D^+D^-] - ED^+ED^- \\
E[D^+D^-] &= E_{\hat{W}^-}[E_{W^-}[E[D^+D^-]|W^-]|\hat{W}^-] \\
&= E_{\hat{W}^-}[E_{W^-}[W^+W^-|W^-]|\hat{W}^-] \\
&= E_{\hat{W}^-}[E_{W^-}[(ad(W^-)^s + a(1-d)(\hat{W}^-)^s)(W^-)^s|W^-]|\hat{W}^-] \\
&= adE(W^-)^{s+1} + a(1-d)E(\hat{W}^-)^s EW^-
\end{aligned}
$$



(a) Expected degree

(b) Alpha

(c) Beta

Figure 1: Parameter effect on degree correlation (α, β and Expected degree)

Finally,

$$\rho = corr(D^+, D^-) = \frac{Cov(D^+, D^-)}{\sqrt{Var(D^+)Var(D^-)}}$$

The complete formula of ρ is complex, so we carry out numerical simulations to find out how do each parameter (α, β, d and expected degree) affect the degree correlation.

According to Figure 1, we could see as α, β, d or expected degree increases independently, the degree correlation increases.

3.1.4 Model Generation

We denote E the expected degree of nodes, and n the graph size, that is, the total number of nodes in a graph, in the following context. In DCM generation, we only need to fix the value of α, β, d, E, and n to determine the model.

1. Degree Sequence Generation
   Let $P(W^- < x) = 1 - (\frac{x}{c})^{-\beta}$, thus the CDF of W^- is

$$z = F(x) = 1 - P(W^- > x) = 1 - (\frac{x}{c})^{-\beta}$$

The inverse function is

$$x = F^{-1}(z) = c(1-z)^{-\frac{1}{\beta}}$$

So the steps are:

Step 1: Generate W^- by

$$U \sim \text{Uniform}(0,1), W^{-1} = F^{-1}(U) = c(1-U)^{-\frac{1}{\beta}}$$

Step 2: Create an independent copy of W^-, denoted by $\hat{W}^-$

Step 3: Generate W^+ by

$$W^+ = ad(W^-)^S + a(1-d)(W^-)^s$$

Step 4: Independently generate $(D^+, D^-) = (Poisson(W^+), Poisson(W^-))$. Apply the sequence-modification algorithm in [1] (a brief description in sequence-modification algorithm in Appendix) to make the sum of in-degree sequence equal to the sum of out-degree sequence.

2. DCM generation (with Erased algorithm)

```
import networkx as nx

 # generate the multigraph
 dcm = nx.directed_configuration_model(self.d_in, self.d_out)

 # remove parallel edges
 dcm = nx.DiGraph(dcm)

 # remove self-loops
 dcm.remove_edges_from(dcm.selfloop_edges())
```

3.2 Real-World Graph

We use *Wikipedia vote network* in the test of real-world graph.
Here is a brief introduction to Wiki-Vote network: 'Wikipedia vote network is a directed graph reflecting the administrator election vote in the Wikipedia community. The graph data is derived from Stanford Large Network Dataset Collection. A small part of Wikipedia contributors are administrators, who are users with access to additional technical features that aid in maintenance. In order for a user to become an administrator a Request for adminship (RfA) is issued and the Wikipedia community via a public discussion or a vote decides who to promote to adminship. The network contains all the Wikipedia voting data from the inception of Wikipedia till January 2008. Nodes in the network represent wikipedia users and a directed edge from node i to node j represents that user i voted on user j.'
A full description and network data is in [12]

4 Part 1: Relationship between Page Rank and Betweeness Centrality

In this part, we want to find out the relationship between page rank and betweenness centrality, especially how it changes as the degree correlation changes. Specifically, by fixing the values of power-law index (α and β), we adjust the value of d from 0 to 1 to vary the degree correlation, and then observe how the relationship between page rank and betweenness centrality changes.

4.1 Method Description

In this part we introduce two measures of correlation between PageRank and Betweeness centrality.

Table 1: Case1 ($\alpha = \beta$): statistics of rank correlation over 20 samples

d	computed degree correlation	average of rank correlation	variance of rank correlation
0	0	0.668137886	7.28353E-05
0.2	0.2047114	0.711738476	5.53925E-05
0.4	0.455694377	0.754644549	4.67439E-05
0.6	0.683541566	0.786671926	0.000186392
0.8	0.818845602	0.816479938	9.33524E-05
1.0	0.869565217	0.8455353	0.000119224

4.1.1 Statistical approach: Spearman's Rank Correlation

'In statistics, a rank correlation is any of several statistics that measure an ordinal associationthe relationship between rankings of different ordinal variables or different rankings of the same variable, where a "ranking" is the assignment of the labels "first", "second", "third", etc. to different observations of a particular variable. A rank correlation coefficient measures the degree of similarity between two rankings, and can be used to assess the significance of the relation between them.' [11]

We use Spearman's test to measure the rank correlation between page rank and betweenness centrality.

4.1.2 Visual approach: Ranking Plot

Make a scatterplot of Page Rank and Betweenness centrality of top k nodes in decreasing order of Page Rank(Betweenness centrality). We expect to see a downward curve of PR (BC) values, and conclude that if the scatter plot of BC (PR) is less scattered and more downward-bahaved, the correlation between Page Rank and Betweenness Centrality is more obvious.

4.2 Experiment and Analysis

In this part, we fix the values of α, β and expected degree and vary the value of d from 0 to 1 to change the degree correlation, and observe how rank correlation changes. For each model we repeat the simulation for 20 times independently to obtain the sample correlation between PR and BC.

4.2.1 Case1: $\alpha = \beta$

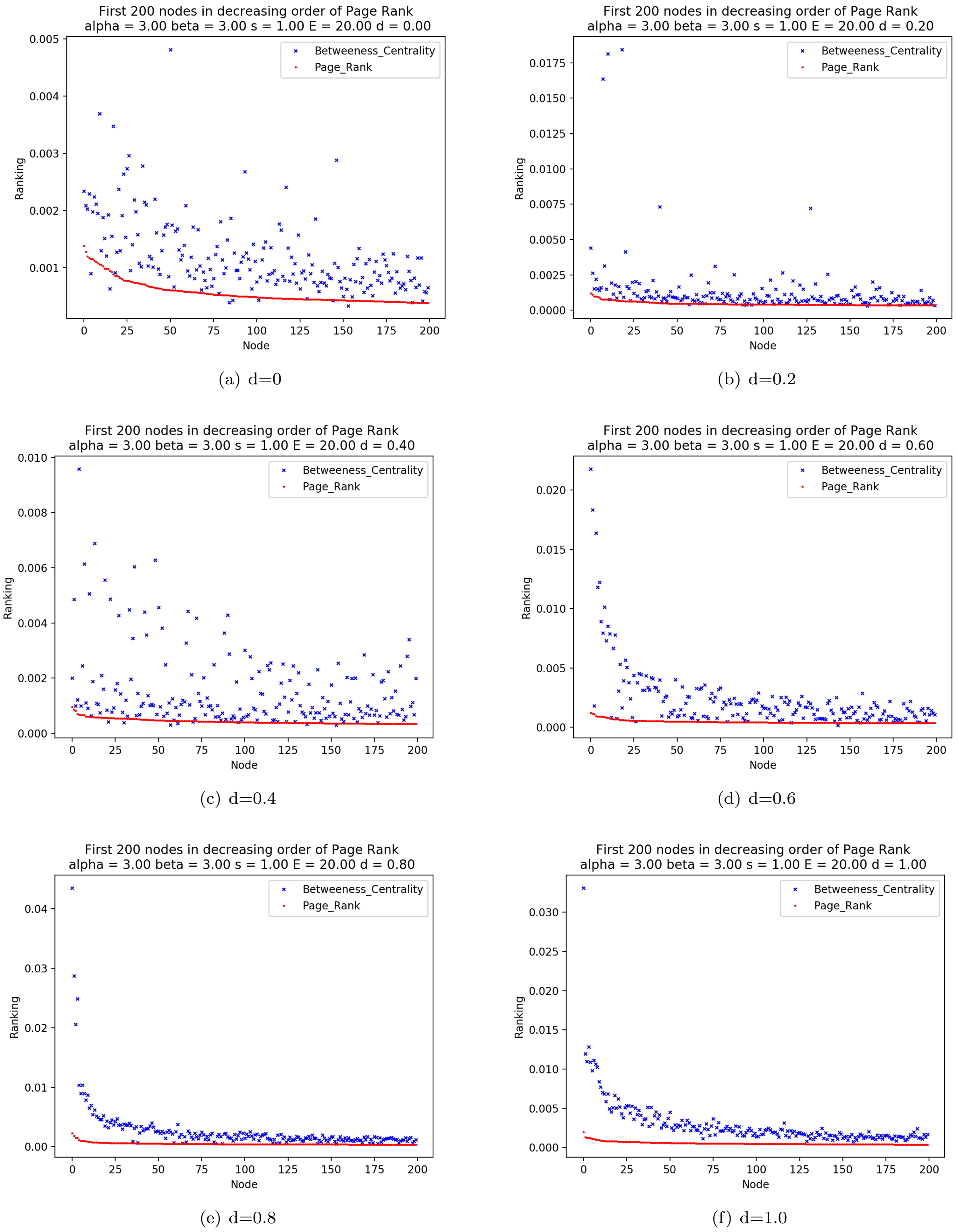
In the simulation, fix $\alpha = \beta = 3, E = 20, n = 5000$ and let d vary from 0 to 1 with step size 0.2, to make the degree correlation vary from 0 to 0.943. The simulation results are given in Table 1, Figure 3, Figure 2, 4, and the detailed information of the models are given in excel files.

1. Ranking Plot
   From Figure 2, we plot PageRank and Betweenness Centrality values of the top 200 nodes in decreasing order of Page Rank, so that the PageRank plot (red color) must be a downward curve. We could see that as d increases (thereby degree correlation increases), dots identifying Betweenness Centrality become less scattered, and behave more like a downward trend as does Page Rank curve. We could observe a similar trend in Figure 3 as well.

2. Rank Correlation

   (a) From Table 1, we could see that the sample variance of rank correlation is close to 0, which suggests that the sample average of rank correlation is representative.

   (b) From Figure 4, we could see as d increases, degree correlation increases, and rank correlation increases.



(a) d=0

(b) d=0.2

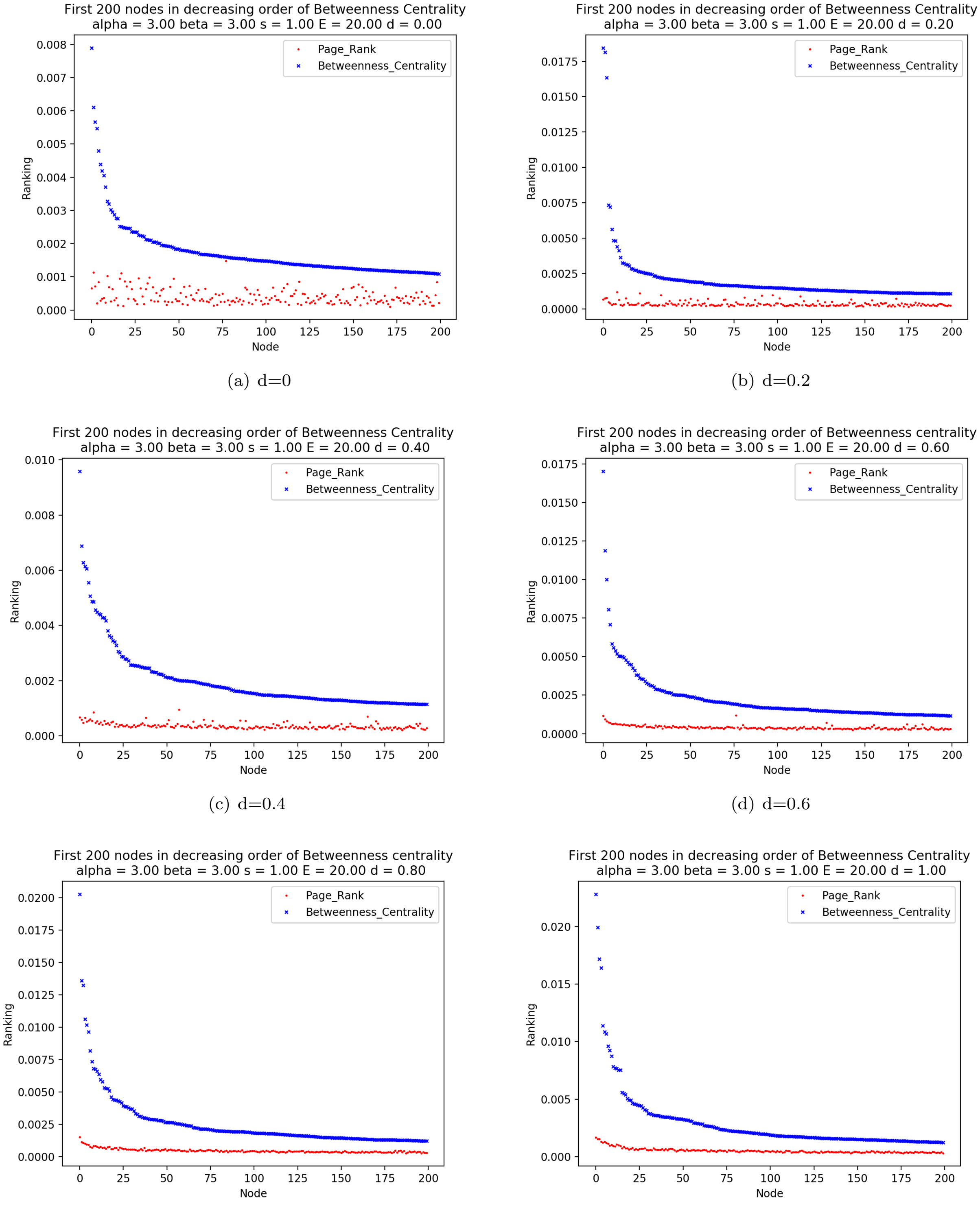
(c) d=0.4

(d) d=0.6

(e) d=0.8

(f) d=1.0

Figure 2: Case1 ($\alpha = \beta$): Page Rank and Betweeness Centrality values of First 200 nodes **in decreasing order of Page Rank** with increasing d values from 0 to 1



(a) d=0

(b) d=0.2

(c) d=0.4

(d) d=0.6

(e) d=0.8

(f) d=1.0

Figure 3: Case1($\alpha = \beta$): Page Rank and Betweenness Centrality values of First 200 nodes **in decreasing order of Betweenness Centrality** with increasing d values from 0 to 1

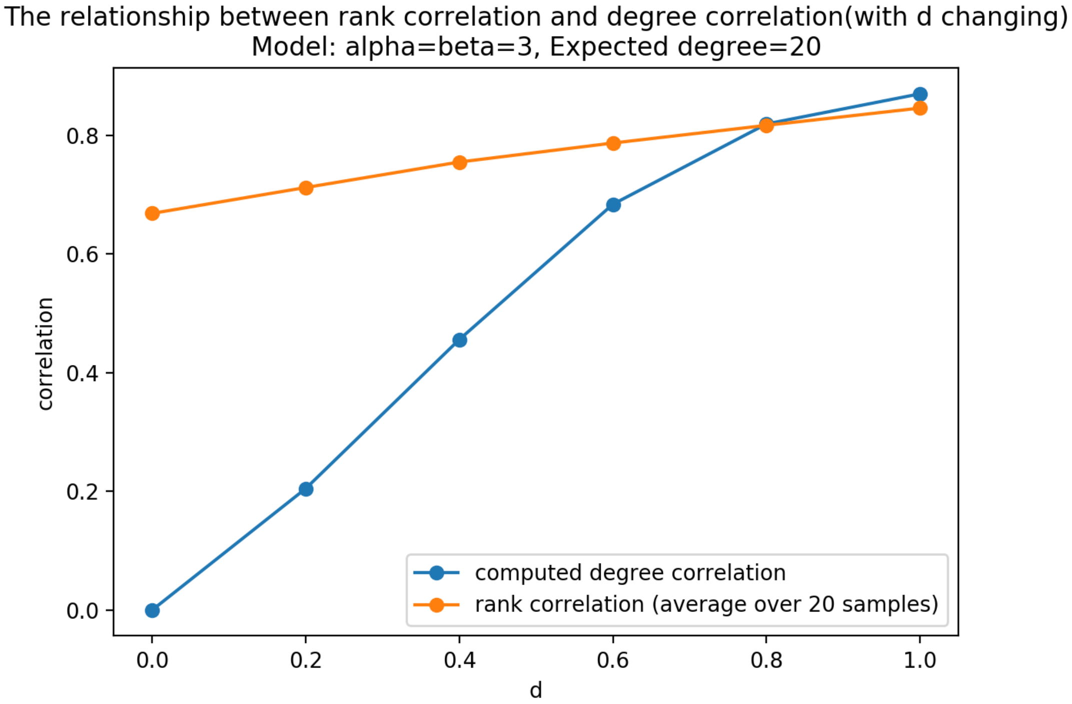
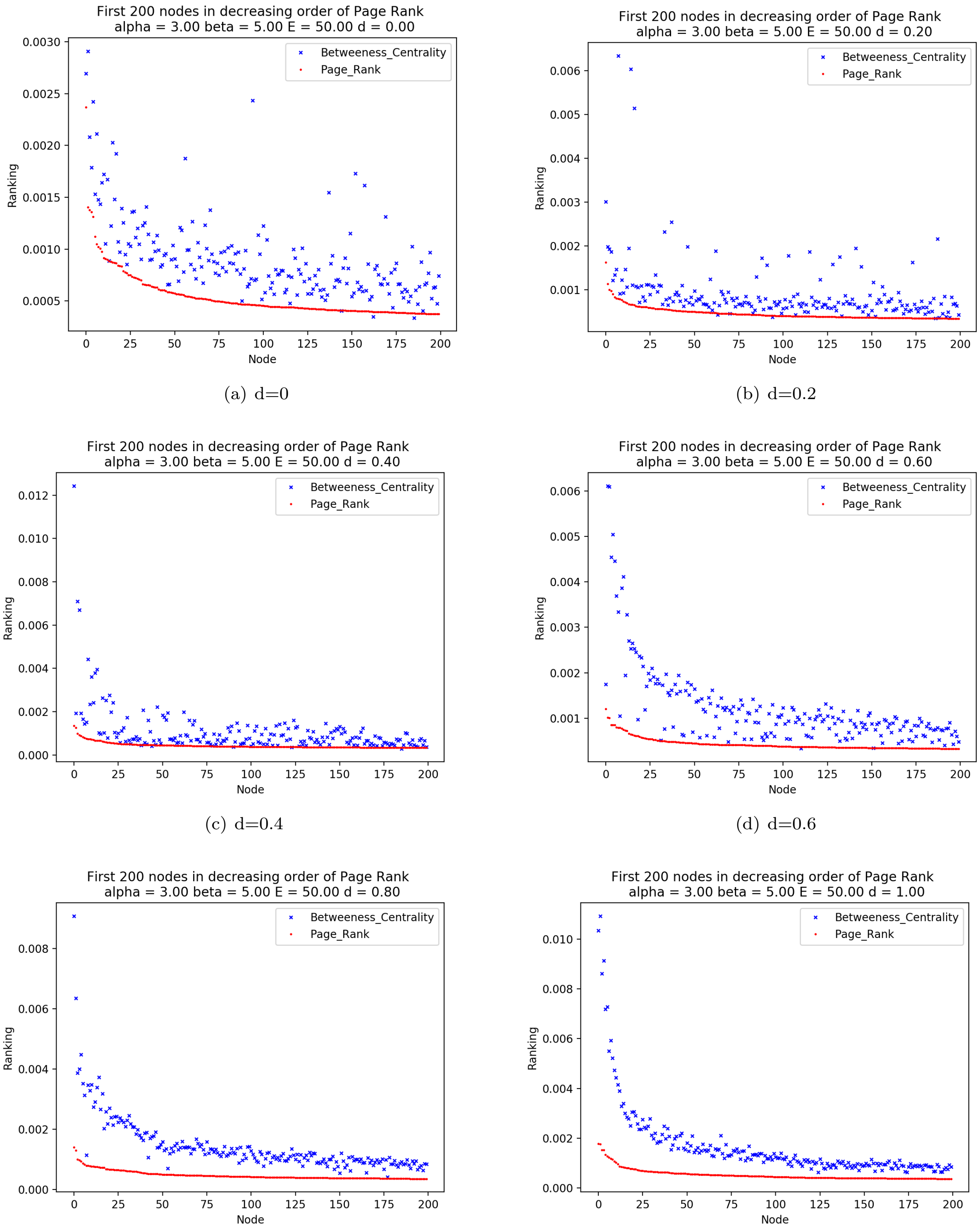


Figure 4: Case1($\alpha = \beta$): relationship between rank correlation and degree correlation

4.2.2 Case2: $\alpha \neq \beta$

In this simulation, fix $\alpha = 3, \beta = 5$, and expected degree $E = 50$, graph size $n = 5000$ and let d vary from 0 to 1, with step size 2, to make the degree correlation vary from 0 to 0.816. The simulation results are given in Table 2, Figure 6, 5,7, and the detailed information of the models are given in excel files.

1. Ranking Plot
   From Figure 5, we plot Page Rank and Betweenness Centrality values of the top 200 nodes in decreasing order of Page Rank, so that the Page Rank plot (red color) is a downward curve. We could see that as d increases (thereby degree correlation increases), scatterplot of Betweenness Centrality becomes less scattered, and behaves more like a downward trend as does Page Rank curve. We could observe a similar trend in Figure 6 as well.



(a) d=0

(b) d=0.2

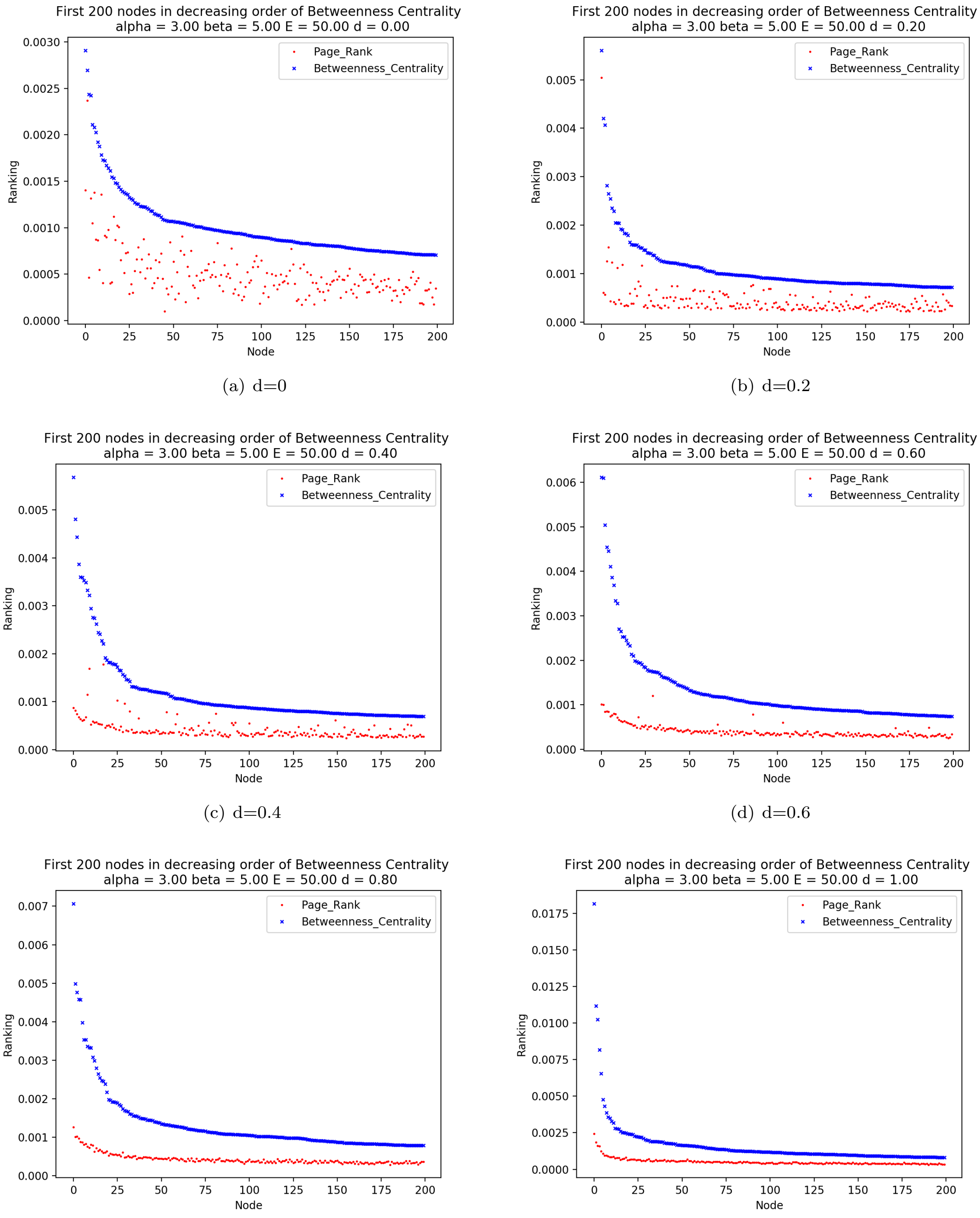
(c) d=0.4

(d) d=0.6

(e) d=0.8

(f) d=1.0

Figure 5: Case2 ($\alpha \neq \beta$): Page Rank and Betweenness Centrality values of First 200 nodes **in decreasing order of Page Rank** with increasing d values from 0 to 1



(a) d=0

(b) d=0.2

(c) d=0.4

(d) d=0.6

(e) d=0.8

(f) d=1.0

Figure 6: Case2 ($\alpha \neq \beta$): Page Rank and Betweenness Centrality values of First 200 nodes **in decreasing order of Betweenness Centrality** with increasing d values from 0 to 1

Table 2: Case2 ($\alpha \neq \beta$): statistics of rank correlation over 20 samples

d	computed degree correlation	average of rank correlation	variance of rank correlation
0	0	0.770526926	0.000107614
0.2	0.195411751	0.820663736	3.15524E-05
0.4	0.441451031	0.856461981	2.05965E-05
0.6	0.662176546	0.880852803	1.74735E-05
0.8	0.781647004	0.899151153	1.21308E-05
1.0	0.816363555	0.915141277	1.365E-05

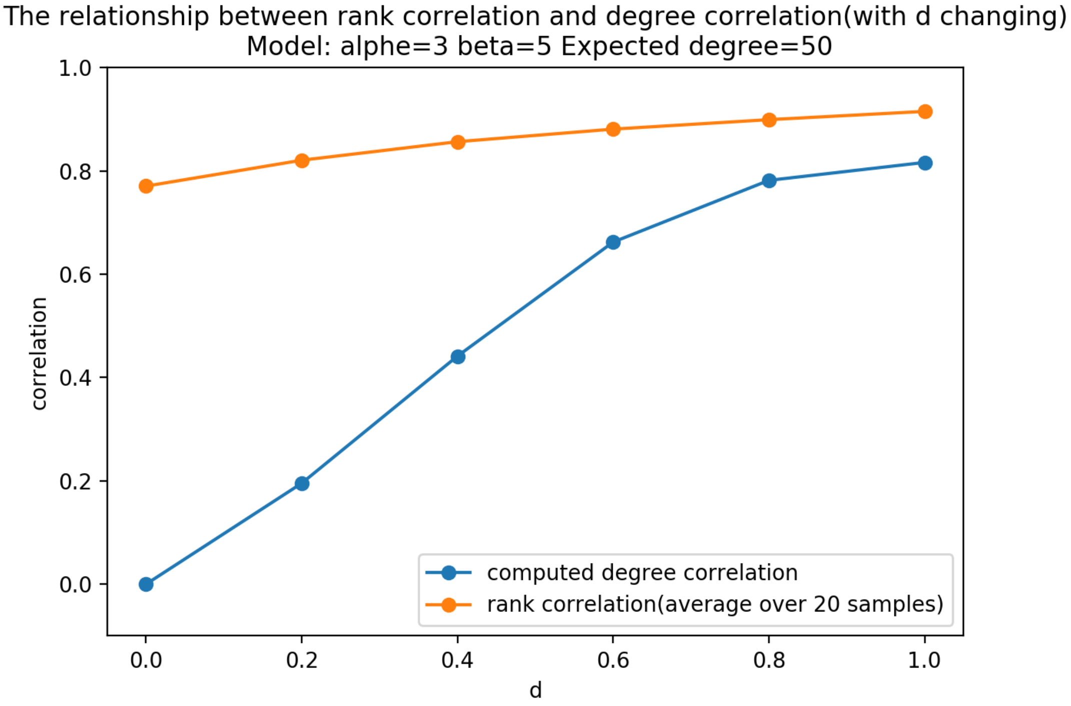


Figure 7: Case2 ($\alpha \neq \beta$): relationship between rank correlation and degree correlation

2. Rank Correlation

   (a) From Table 2, we could see that the sample variance of rank correlation is close to 0, which suggests that the sample average of rank correlation is representative.

   (b) From Figure 7, we could see as d increases, degree correlation increases, and rank correlation increases.

4.3 Summary

1. The noise in average sample rank correlation is very small.

2. As the degree correlation increases (by adjusting the value of d), the correlation between Page Rank and Betweenness Centrality increases.

5 Part 2: Comparison among the effect of Measures on Average Shortest Path Length

In this section, we want to compare among different ranking algorithms / measures of importance (PR, BC, total degree, in-degree, out-degree) with respect to graph connectivity efficiency. For a graph G, we choose the Average Shortest Path Length (ASPL) as the indicator of its connectivity efficiency.

5.1 Method Description

In procedure, we design the 'node-elimination experiment', which is to consecutively eliminate the individual node based on different measures and compute the corresponding ASPL, thus deriving the effect from higher ranking nodes. Note that each time we eliminate a node, we use the remaining graph's strongly connected component(SCC) to calculate the new ASPL, since the compuation of ASPL requires the graph to be connected. After consecutively eliminating m nodes, we could obtain a numeric array of ASPL of length m with respect to one specific ranking algorithm's measure. Independently conducting the procedure w.r.t. each ranking measure, we can make the plot as in Figure 8: the ranking of the lines from top to down suggest the ranking of the effect on ASPL. By observing the trends in the plots, we expect to find out the changes in ASPL are mostly subject to which measure.

Due to the randomness of the graph, there is noise in the simulated graph's ASPL. Therefore, we repeatedly conduct the model simulation and take the average of each simulation's shortest to smooth out the noise.

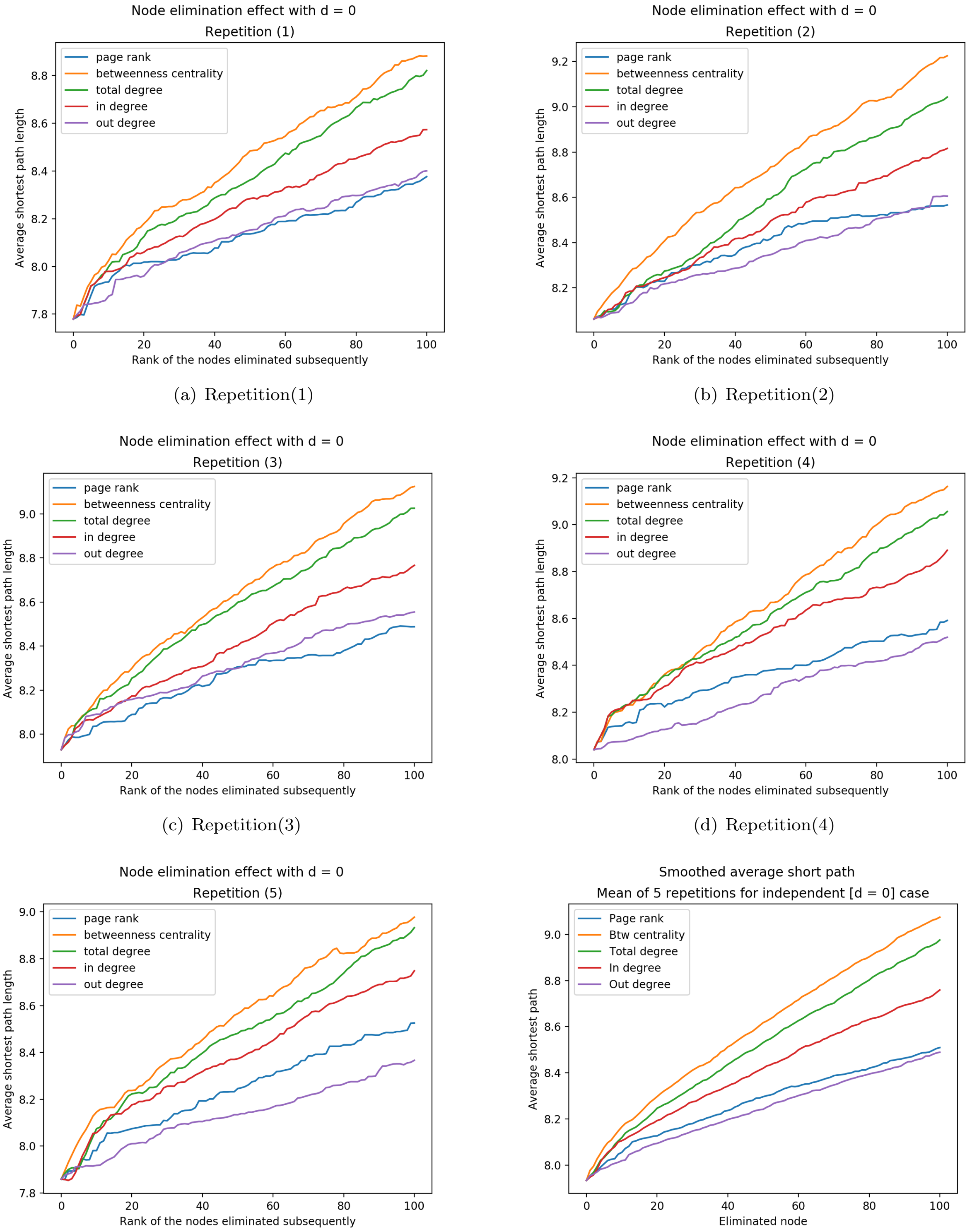
5.2 Experiment and Analysis

5.2.1 Theoretical Models

As we do in Part 1 to vary the degree correlation, we let d range from 0 to 1 with step size 0.25. And the other parameters to fix the model are: $\alpha = 2.1$, $\beta = 3$, the expected degree E = 3, the node size of graph n = 5000. We consecutively eliminate 100 nodes and independently repeat the procedure for 5 times and then compute the average. The results of repetitions in the case $d = 0$ are given in 8. The results of six diferent dependency levels are given in 9 and data in csv files.

1. Noises in ASPL

   The noise can be clearly observed in the first 5 plots ((a) to (e)) in Figure 8. The PR curve and out-degree curve even switch the ordering in different experiments, which suggests the vacillation of outcomes. Plot (f) in Figure 8 is the average effect of the foregoing five independent repetitions, which smooths out the noise. Moreover, it denotes that the changes in ASPL is most sensitive to BC, which makes sense since BC is a measure of centrality. However, it seems that Page Rank fails to capture the ASPL characteristics.



(a) Repetition(1)

(b) Repetition(2)

(c) Repetition(3)

(d) Repetition(4)

(e) Repetition(5)

(f) Average over 5 repetitions)

Figure 8: Repetitive model simulation and the average ranking node elimination effect on ASPL (Models: $\alpha = 2.1, \beta = 3$, Expected degree = 3, d=0, degree correlation=0 and $n = 5000$)

2. Effect of Measures on ASPL in Different dependence levels

In order to check the ASPL change in different dependence (degree correlation) level, we assign d with value from five levels: 0, 0.25, 0.5, 0.75 and 1. And to smooth out the noise, in each dependence level we simulate models by 5 independent replications and then plot the average. In addition, we add the perfectly-correlated case, that is in-degree sequence is identical to out-degree sequence, in which case both in and out degree sequence have power law distribution with index *alpha*.

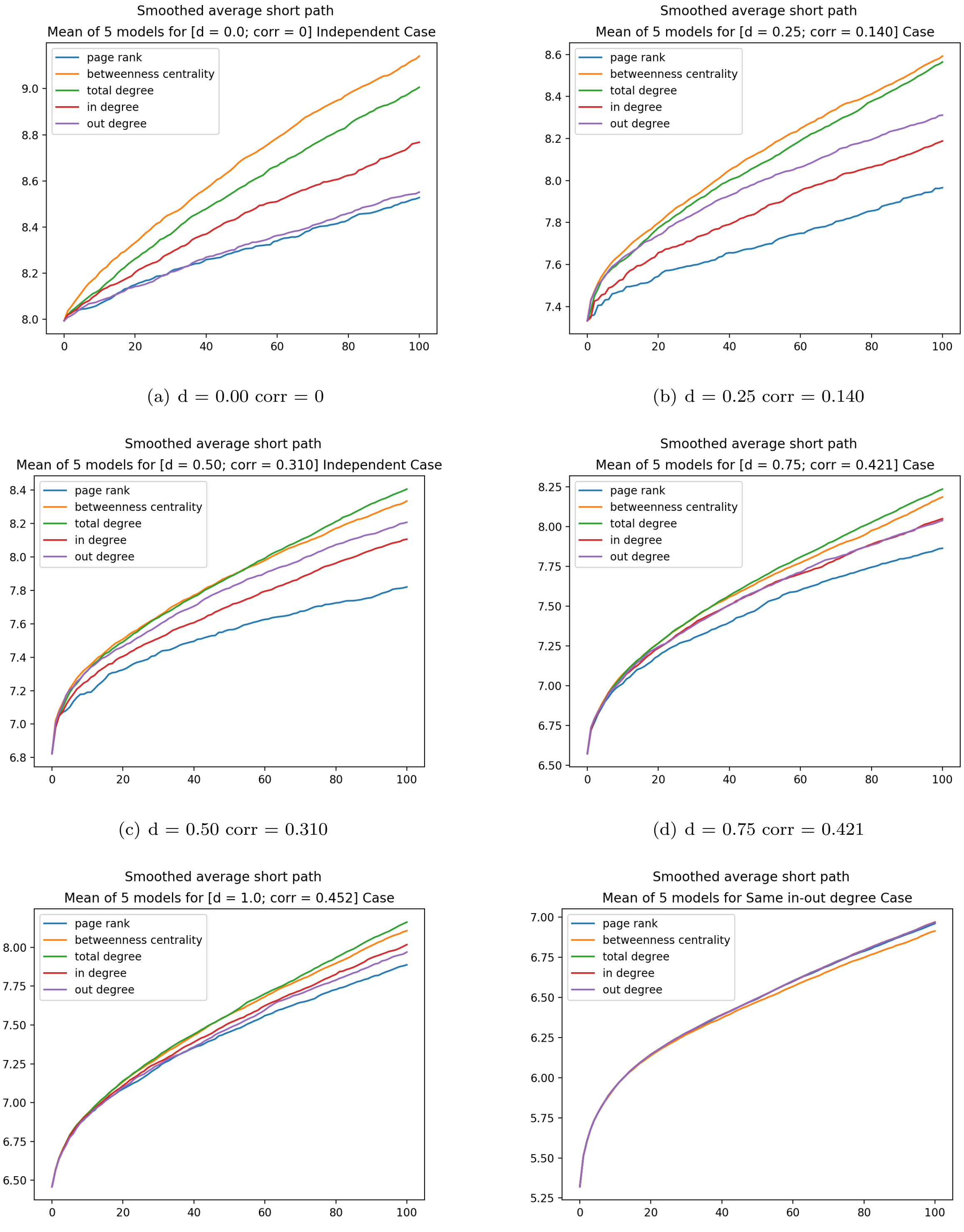
From Figure 9, we have following observations:

(a) As degree correlation increases, the five lines are getting closer, which suggests the effect of each measure is more similar.

(b) From (a) to (e), the ranking of the measures on by effects on ASPL (judged by the order of lines from top to bottom) is: BC/total degree – in-degree/out-degree – PageRank.

(c) In (f), that is the perfectly-correlated case, all measures except BC has similar top effect.

(d) As degree correlation increases, BC's dominance in effects on ASPL declines, while by contrast total degree increases its influence on ASPL.



(a) d = 0.00 corr = 0

(b) d = 0.25 corr = 0.140

(c) d = 0.50 corr = 0.310

(d) d = 0.75 corr = 0.421

(e) d = 1.00 corr = 0.452

(f) corr = 1(perfectly correlated, in degree sequence = out degree sequence)

Figure 9: Comparison of Effects on ASPL in Different Degree Correlation Levels (Models: $\alpha = 2.1, \beta = 3$, Expected degree = 3 and $n = 5000$)

5.2.2 Real World Graph: Wikipedia Vote Network

To testify the numerical study, we apply the experiment to real world data.

1. Introduction of Wiki-Vote Network

   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.

   The statistics of the graph data is shown in Table 3.

   Table 3: Wikipedia vote network statistics

Feature name	Value
Nodes size	7115
Edges	103689
Expected degree	14.57
Numer of nodes in the largest SCC	1300
In-out degree correlation	0.317
Average clustering coefficient	0.1409
Number of triangles	608389
Diameter(longest shortest path)	7

2. Fitting Wiki-Vote Network into Theoretical Model

   To determine whether Wiki-vote network has scale-free property (power law degree distribution), and what are the indices if it has, we make the log-log plot of the tail-distribution of in-degree and out-degree sequence, and fit the tail. The fitting result is shown in Figure 10 :$alpha = 3.5, \beta = 2.5$, since we have $E = 14.57$, in-out degree correlation $= 0.317$, we can substitute the values into the degree correlation formula to get $d = 0.339$.

3. Node Elimination Effect on Wiki-vote Network

   We consecutively eliminate 200 nodes from Wiki-vote network. The node elimination effect on the graph's average shortest path length is shown in Figure 11, and the output data is here.

   Figure 11 shows that PageRank is the worst in measuring the significance of ASPL, while both betweenness centrality and total degree perform well with the former slightly better. The trend and relative ranking in Figure 11 is consistent with previous simulation results given the similar degree dependence level.

4. Node Elimination Effect on Simulated Model of Wiki-vote Network

   We further generate the theoretical model to fit Wiki-vote network. From Table 3 and the result of tail-distribution fitting, we have parameters: $\alpha = 3.5$, $\beta = 2.5$, E = 14.57, d = 0.339 and n = 7115. We consecutively remove 200 nodes and obtain the results in Figure 12 and the output data files.

   Note that there exists inflexible difference between fitting model and real graph. Specifically the clustering coefficient, the real data has average clustering coefficients of 0.1409, while the simulation graph is just in the level of 0.005. This time, we repeat the model simulation for 10 times and measure the effect upon ASPL for consecutively 200 nodes. As before, we take the average to reduce the noise.

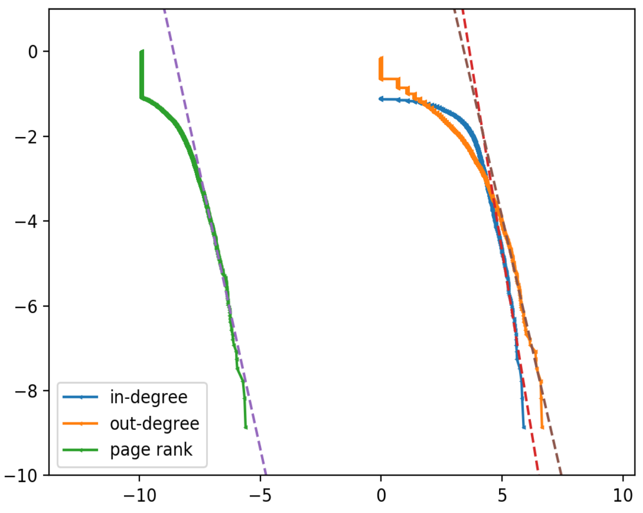


Figure 10: log-log plot of tail distribution of Wiki-vote network

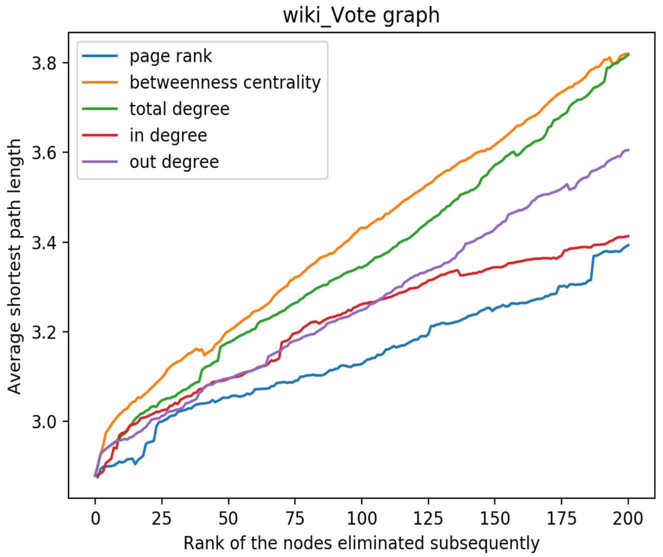


Figure 11: Node elimination effect on ASPL of Wiki-vote network.

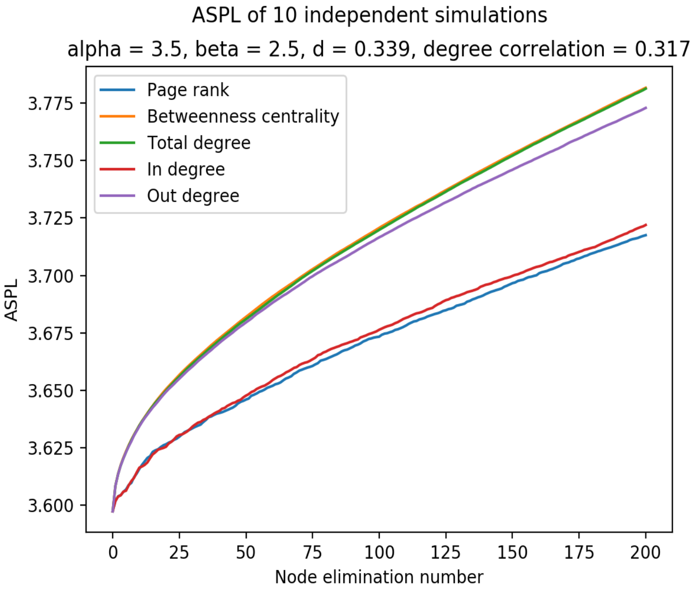


Figure 12: Average node elimination effect on ASPL of Wiki-Vote simulated models ($\alpha = 3.5, \beta = 2.5$, Expected degree $= 14.57$, $d = 0.339$ and $n = 7115$)

From Figure 12, we could see the result is generally similar with the real data in Figure 11 with same ranking of measures' performance: BC–total degree–out-degree-in-degree–PR.

Nevertheless, there is some difference:

(a) In the case of simulation model, the effect of total degree is almost the same as that of betweenness centrality, whereas in Wiki-vote network betweenness is better.

(b) In the case of simulation model, the effect of in-degree is close to, but slightly better than, PageRank, while in Wiki-vote network the difference is bigger.

(c) Wiki-vote network has larger range of ASPL which from 2.8 to 3.8, while in simulation model it is from 3.6 to 3.8. The potential explanation may be that the simulated graph presents the average situation, which reflecting the expected effect, in contrast, the real graph is a sample realization, it has more noise.

5.3 Summary

1. In general, BC and total degree mostly affect ASPL, while PR is the worst indicator, with in-degree and out-degree in between.
2. Dependence level (degree correlation) affects the relative ranking of the effects of each measures. Specifically, dependence is negatively correlated with the betweenness centrality performance, while total degree maintains an effective measure; And the more correlated is the in- and out-degree sequence, the smaller the measure differences in ASPL.

6 Conclusions

We study the relationship between PR and BC, and their effects on ASPL, an indicator of graph's connectivity efficiency, together with degree measures(total degree, in-degree, out-degree). We numerically generate the directed configuration model as fundamental tools and processed the study respectively into two parts.

In the first section, we show that the degree dependence positively affects the correlation between PR and BC. Controlling power-law indices and other scaling factors, we vary degree-degree correlation to conduct experiments. After repeatedly and independently sampling, we smoothe out the noise by compuitng the average, and conclude that degree dependence could increase the ranking similarity of PR and BC.

In the second study, we show that, in general, among PR, BC, degree measures(total degree, in-degree and out-degree), the changes in ASPL are most sensitive to PR and total degree, and then in-degree and out-degree have medium effects, with PR the worst. Furthermore, investigation into effects of degree correlation on the experiments shows that given higher degree, the sensitivity of changes in ASPL with respect to BC is declining, while total degree is still in a dominant place. In addition, the difference among performance of five measures is narrowed with degree correlation increased. As a supplement, a real directed graph–Wiki-vote network–is studied with the same method, the results of which are consistent to those of simulated random graphs. In general, the stable performance of total degree suggests that it is the most appropriate choice, among the five measures, in the evaluation of ASPL

7 Appendix

1. Proof of Dependencies of Parameters:
   From the relationship between W^+ and W^-:

$$W^+ = a \cdot d \cdot {W^-}^s + a \cdot (1-d) \cdot \hat{W^-}^s$$

where $\hat{W^-}$ is the independent copy of W^-. Applying the double integral, the probability of W^+ is thus the following:

$$\begin{aligned} P(W^+ > x) &= P(d \cdot {W^-}^s + (1-d) \cdot \hat{W^-}^s > \frac{x}{a}) \\ &= (\frac{c_1}{c})^{-\beta} + (\frac{c_2}{c})^{-\beta} - (\frac{c_1 c_2}{c^2})^{-\beta} \end{aligned}$$

where c_1 is the integral boundary value for W^-: $c_1 = \frac{1}{d}^{\frac{1}{s}} \cdot (\frac{x}{a} - (1-d) \cdot c^s)^{\frac{1}{s}}$; and c_2 for $\hat{W^-}$: $c_2 = \frac{1}{1-d}^{\frac{1}{s}} \cdot (\frac{x}{a} - d \cdot c^s)^{\frac{1}{s}}$. Hence the power law distribution:

$$\lim_{x \to +\infty} \frac{(\frac{c_1}{c})^{-\beta} + (\frac{c_2}{c})^{-\beta} - (\frac{c_1 c_2}{c^2})^{-\beta}}{x^{-\alpha}} = Constant$$

To satisfy the power law requirement, it suffices that:

$$s = \frac{\beta}{\alpha}$$

By definition, the lower bound for W^+ is b and the lower bound for W^- and $\hat{W}^-$ is c, so we have:

$$b = a \cdot d \cdot c^s + a \cdot (1-d) \cdot c^s = a \cdot c^s$$

Additionally, as the mean of in-degree should be equal to that of out-degree, we have:

$$\begin{aligned} E(W^+) = E(W^-) &\Rightarrow a \cdot E({W^-}^s) = E({W^-}^s) \\ &\Rightarrow a \cdot \frac{\beta}{\beta - s} \cdot c^s = \frac{\beta}{\beta - 1} \cdot c \end{aligned}$$

Specifically in consideration of the case where s = 1, given that case in-degree and out-degree are supposed to be equal, thus identical W^+ and W^-, a should also be 1.

2. Sequence-modification Algorithm

   Given the in-out degree distribution from realized W^+ and W^-, we generate the modified in-out degree sequence satisfying equal sum of in-out degree sequence while approximately maintaining the previous distribution.
   In the first place, we derive the constant value κ.

$$\kappa = min\{1 - \alpha^{-1}, 1 - \beta^{-1}, \frac{1}{2}\}$$

Step 1: Fix $0 < \delta_0 < \kappa$, specifically choose $\delta_0 = 0.995 * \kappa$

Step 2: Sample an i.i.d. sequence $\{\gamma_1, \cdots, \gamma_n\}$ from in-degree distribution the poisson of W^+; let $\Gamma_n = \sum_{i=1}^n \gamma_i$

Step 3: Sample an i.i.d. sequence $\{\xi_1, \cdots, \xi_n\}$ from out-degree distribution the poisson of W^-; let $\Xi_n = \sum_{i=1}^n \xi_i$

Step 4: Define $\Delta_n = \Gamma_n - \Xi_n$, If $|\Delta_n| \leq n^{1-\kappa+\delta_0}$ proceed to step 5; otherwise repeat from step 2.

3. Data and Codes Avalibility: Project's GitHub Page

   This includes the full Python code of this project, from building the graph, testing and analyzing are in the project's GitHub page, weekly reports, a final report, an output data folder and a README file describing the data files.

References

[1] Ningyuan Chen and Mariana Olvera-Cravioto, *Directed Random Graphs with Given Degree Distributions July 12, 2012* https://arxiv.org/pdf/1207.2475.pdf

[2] Project's GitHub page https://github.com/leahwu/Go_graphs

[3] networkx package https://networkx.github.io/documentation/networkx-1.11/ Released Jan 30, 2016

[4] directed_configuraiton_model https://networkx.github.io/documentation/networkx-1.11/reference/generated/networkx.generators.degree_seq.directed_configuration_model.html?highlight=directed%20configuration%20model#networkx.generators.degree_seq.directed_configuration_model

[5] A. Langville and C. Meyer, A survey of eigenvector methods of web information retrieval. http://citeseer.ist.psu.edu/713792.html

[6] Page, Lawrence; Brin, Sergey; Motwani, Rajeev and Winograd, Terry, The PageRank citation ranking: Bringing order to the Web. 1999 http://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en&doc=1999-66&format=pdf

[7] Ulrik Brandes: A Faster Algorithm for Betweenness Centrality. Journal of Mathematical Sociology 25(2):163-177, 2001. http://www.inf.uni-konstanz.de/algo/publications/b-fabc-01.pdf

[8] (1, 2) Ulrik Brandes: On Variants of Shortest-Path Betweenness Centrality and their Generic Computation. Social Networks 30(2):136-145, 2008. http://www.inf.uni-konstanz.de/algo/publications/b-vspbc-08.pdf

[9] Ulrik Brandes and Christian Pich: Centrality Estimation in Large Networks. International Journal of Bifurcation and Chaos 17(7):2303-2318, 2007. http://www.inf.uni-konstanz.de/algo/publications/bp-celn-06.pdf

[10] Linton C. Freeman: A set of measures of centrality based on betweenness. Sociometry 40: 3541, 1977 http://moreno.ss.uci.edu/23.pdf

[11] Spearman's ranking coefficient https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

[12] Wiki-vote website and data https://snap.stanford.edu/data/wiki-Vote.html