Random Graph Standard Network Metrics Distributions in ORA*

Kathleen M. Carley and Eunice J. Kim

March, 2008
CMU-ISR-08-103

Institute for Software Research
Carnegie Mellon University
School of Computer Science
Pittsburgh, PA 15213

Center for the Computational Analysis of Social and Organizational Systems
CASOS technical report.

This work was supported in part by the National Science Foundation IGERT grant 9972762 and the Office of Naval Research N00014-06-1-0104 and N0001406-1-0921 and by CASOS, the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the National Science Foundation, the Office of Naval Research, or the U.S. government.

Keywords: random network, distribution, centrality, path length, clustering coefficient

Abstract

Networks, and the nodes within them, are often characterized using a series of metrics. Illustrative graph level metrics are the characteristic path length and the clustering co-efficient. Illustrative node level metrics are degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality. A key issue in using these metrics is how to interpret the values; e.g., is a degree centrality of .2 high? With normalized values, we now that these metrics go between 0 and 1, and while 0 is low and 1 is high, we don't have much other interpretive information. Here we ask, are these values different than what we would expect in a random graph. We report the distributions of these metrics against the behavior of random graphs and we present the 95% most probable range for each of these metrics. We find that a normal distribution well approximating most metrics, for large slightly dense networks, and that the ranges are centered at the expected mean and the endpoints are two (sample) standard deviations apart from the center.

Table of Contents

1 Introduction

In social network analysis, we summarize many features of networks in numerical forms. Some representative measures we use are degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality at the node-level, and we add the characteristic path length and clustering coefficient at the graph-level. Degree centrality shows the connectivity of each node to other nodes in the network; betweenness centrality shows a node's relative position through being an interconnected node of at least one pair of nodes; closeness centrality shows the inverse of the sum of distances from a node to the rest of the nodes in the network; and eigenvector centrality shows a node's importance in the network accounting for the connectedness of its' neighboring nodes. [Wasserman & Faust(1994)]
These network statistics should be interpretive, and we find it best to set a random graph as a benchmark to represent the typical ranges of metrics and compare with any network data. In (this version of) ORA, we report the distributions of above-mentioned graph metrics so that one could check input network data against the behavior of random graphs. We present the 95% most probable range of these metrics. As we find a normal distribution well approximating most metrics of large networks, the ranges are centered at the expected mean and the endpoints are two (sample) standard deviations apart from the center. For centralizations, it is to measure the overall mean of the spread from the maximum centrality value. Hence, we have skewed distribution toward lower values.

2 Data

We simulate random graphs given a fixed number of nodes and density based on the probabilistic formation of links. In a random graph, each link is present independent of other links with a fixed probability p. If there are N nodes in a graph, and each node is connected to an average of k other nodes, then it is easy to show that $p = \frac{k}{N-1}$ which for large N is usually approximated by $\frac{k}{N}$.
We generate 250 random graphs each of size $N = 10, 25, 50, 100, 250, 500,$ and 1000 and each density $d = 0.01, 0.02, 0.05, 0.1, 0.3$, and for each size $N = 10, 25, 50, 100$ we also have 250 random graphs with density $d = 0.5, 0.7,$ and 0.9. We represent a graph G in an $N \times N$ adjacency matrix. In the $G(i, j)$ entry we have either 0 to represent the absence of a link or 1 to represent the presence of a link between the i^{th} node and the j^{th} node. Hence, an undirected graph renders a symmetric (adjacency) matrix, and a directed graph not.
Using ORA, we calculate the standard network metrics of these graphs. From the collection of these values, we derive distributional properties of the network centralities, centralizations, characteristic path length, and clustering coefficient.

3 Methods of Testing Normality of a Distribution

From our data generation, we have 250 values of each graph-level measure, as we created 250 graphs for a given network size and density. We also have 250 times N values of node-level measures, for N is the size of network or the number of nodes in the network.
When we obtain these values, we may plot histograms to glance at the overall shape of a distribution and also Quantile-Quantile plot (so called, Q-Q plot) in order to prove or disprove the normality of the distribution. For Q-Q plots we have the ordered sample on the vertical axis and the normal theoretical quantiles on the horizontal axis. When the data is distributed normally, a match of ordered sample and proposed theoretical normal quantiles forms a straight line. Regardless of a few points at the tails, we see a linearity relationship near the center. For example, in Figure1 the Q-Q plots of characteristic path lengths becomes linear as the density increases. In the bottom-right, when $d = 0.1$, the sample distribution does follow a normal distribution. In the case of bottom-left, when $d = 0.05$, the middle 95% of the distribution (within 2 standard deviations from '0') follows the straight theoretical normal quantile values, so we could well approximate this distribution with a normal, also. However, for the cases of the top two Q-Q plots they are highly non-linear, so we do not use a normal distribution to approximate.

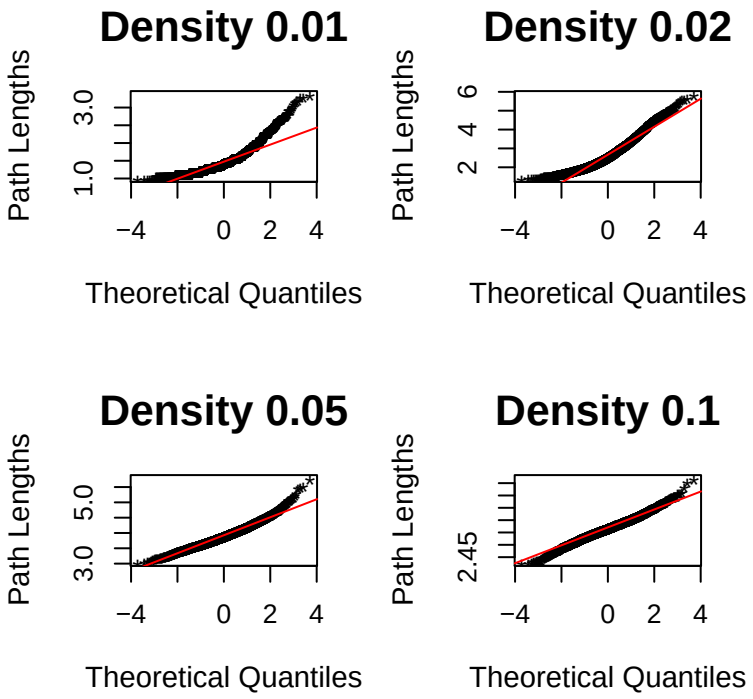


Figure 1: Example of Q-Q plots: checking the normality of characteristic path length distributions when network size is 50 and the density varies from 0.01 up to 0.1.

4 Probable Ranges of Random Digraph Metrics

4.1 Clustering Coefficients

For directed random graphs in our simulated data see Figure 2, the clustering coefficient converges to the true density of the network as the network size grows. Also, we see in Figure 3 the standard deviation decreasing as a function of network size. From our testing, it exhibits an exponential decay in a following form (see Table 1 for specifics):

$$\mu_{cc} = d$$
$$\sigma^2_{cc} = k * exp^{-\frac{N}{\alpha}} + c \qquad \text{where } d = \text{density and } N = \text{graph size.}$$

We test the normality of the data and see them pass when $N \times d \geq 5$. Since we have learned that the mean is independent of the network size, the center of the probable range is $\hat{d}$. The width of the range is $2 \times (\hat{\sigma}_{d,\ N)})$ (sample standard deviation) as the normal distribution has 95% of the area under the curve fall within two standard deviations. Then, we obtain a 95% probable range of the clustering coefficient for a given network size N (at fixed density 0.01, 0.02, 0.05, 0.1, 0.3, and 0.5) as $\hat{d} \pm 2\hat{\sigma}_{cc\ \ (d,\ N)}$. See Table1 for a specific (density, size) pair standard deviation $\hat{\sigma}_{(d,\ N)}$, and see Figure3 for the shape of the sample standard deviation functions.

Table 1: Network (density, size) pair clustering coefficient standard deviations

Density (d),	Sizes (N)	Standard Deviation $(\hat{\sigma}_{(d,N)})$	95% Probable Range
0.01	25-1,000	$9.281 \times 10^{-3} \exp^{-\frac{N}{232.445}}$	$0.01 \pm 2\hat{\sigma}_{(0.01,\ N)}$
0.02	25-200	$2.2226 \times 10^{-2} \exp^{-\frac{N}{98.407}}$	$0.02 \pm 2\hat{\sigma}_{(0.02,\ N)}$
	200-1,000	$1.86 \times 10^{-4} + 1.0645 \times 10^{-2} \exp^{-\frac{N}{150}}$	
0.05	25-100	$5.9281 \times 10^{-2} \exp^{-\frac{N}{36.943}}$	$0.05 \pm 2\hat{\sigma}_{(0.05,\ N)}$
	100-1,000	$1.28 \times 10^{-4} + 1.3094 \times 10^{-2} \exp^{-\frac{N}{100}}$	
0.1	25-250	$7.34 \times 10^{-4} \exp^{-\frac{N}{25}}$	$0.1 \pm 2\hat{\sigma}_{(0.1,\ N)}$
0.3	25-250	$3.26 \times 10^{-4} \exp^{-\frac{N}{25}}$	$0.3 \pm 2\hat{\sigma}_{(0.3,\ N)}$
0.5	25-250	$1.20 \times 10^{-4} \exp^{-\frac{N}{35}}$	$0.5 \pm 2\hat{\sigma}_{(0.5,\ N)}$

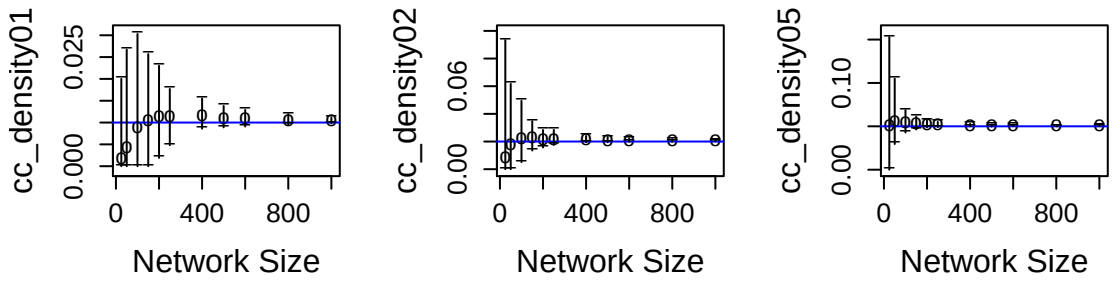


Figure 2: **Convergence of clustering coefficients to a network's true density.** From left to right, each graph shows the random graph density set at 0.01, 0.02, and 0.05 respectively with a blue horizontal line. Means of 100 clustering coefficients for each N are marked with "o", and the range of each distribution is shown by the vertical line. We see the range of clustering coefficients shrinking almost exponentially as N grows.

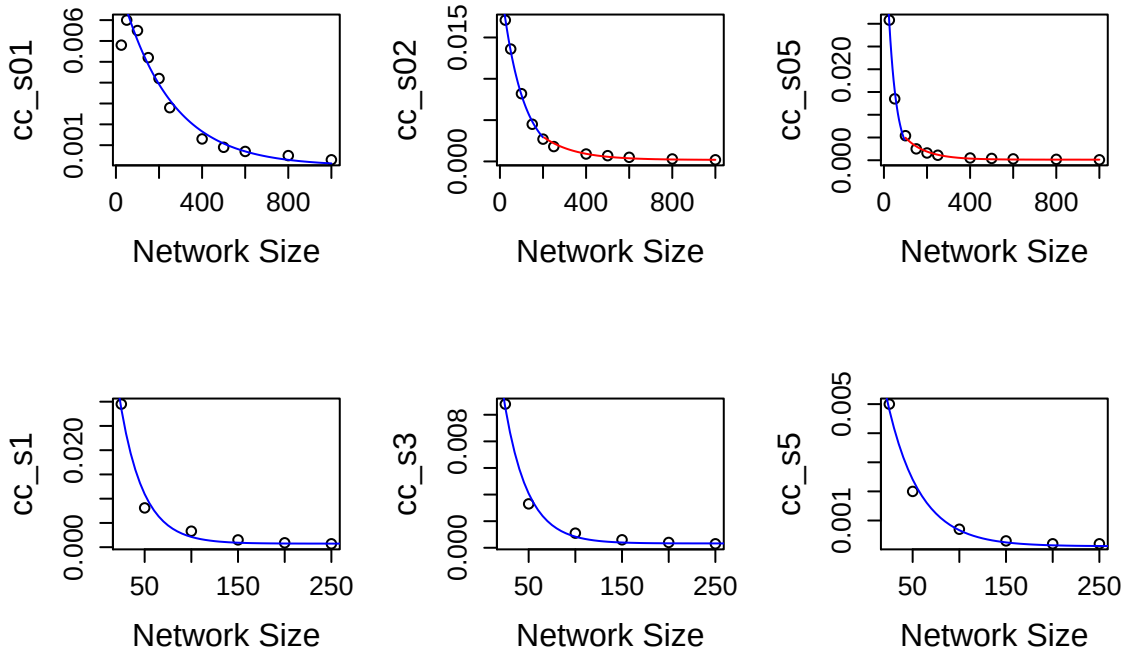


Figure 3: **Interpolation of clustering coefficient standard deviations at fixed densities.** Going clockwise from top-left the densities are d = 0.01, 0.02, 0.05, 0.5, 0.3, and 0.1.

5 Probable Ranges of Undirected Network Graph-level Metrics

5.1 Clustering Coefficient

The clustering coefficient C for the whole network is the average of all nodes' clustering coefficients, C_i for $i \in \{1, \ldots, N\}$, which is defined to be the ratio between the degree (actual number of linked neighbors) and the number of possible links to form a clique with the neighbors. That is, $C_i = \frac{2d_i}{d_i(d_i-1)} = \frac{2}{d_i-1}$ where d_i is the degree of node i and $C = \sum_i \frac{C_i}{N}$. Just as in the case of directed random graphs, for undirected random graphs the clustering coefficient converges to the true density of the network as the network size grows. They behave the same because the directional property of connectedness does not contribute to the count of links. The clustering coefficient's standard deviation again decays exponentially as the network size grows, and when the network size is big ($N > 200$), it remains close to a very small constant (< 0.005).

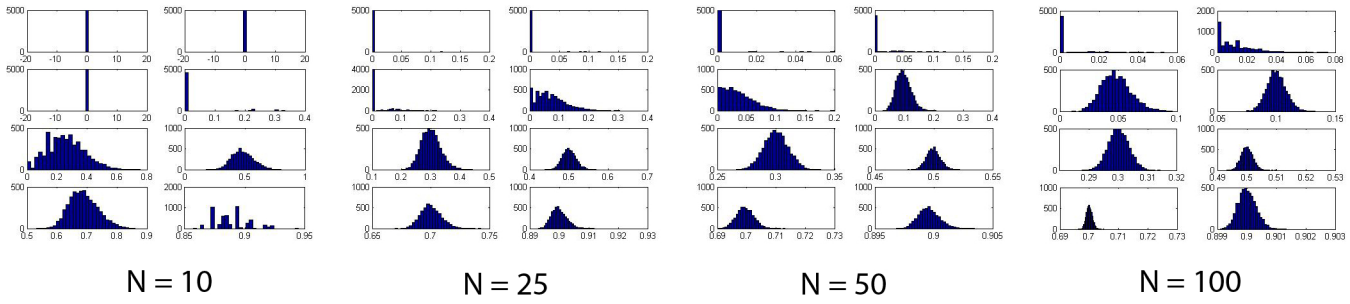


Figure 4: Distributions of clustering coefficients by network size($10 \leq N \leq 100$) and density. For each panel we have density from top-left to right and down 0.01, 0.02(right), 0.05, 0.1(right), 0.3, 0.5(right) and 0.7, 0.9(right).

Except for the case where the network size and density are both small ($N \times d < 5$), the distributions of the clustering coefficients are normal. Hence, the 95% probable intervals of the random network clustering coefficients are as described in Table1.

5.2 Characteristic Path Length (Average Distance)

The characteristic path length L has a lower bound $2d$ where the network density is d [Lovejoy & Loch(2003)]. If we look at Figure 5, for example, when $d = 0.9$, the characteristic path length already piles up around $1.1 = 2 - 0.9$ at $N = 10$; when $d = 0.7$, L is also piled up at 1.3 for networks size $N = 10$ and up; and when $d = 0.5$, it stays around 1.5 starting from a network size $N = 25$. As the random network size grows and/or the density grows, we get a smaller standard deviation of the characteristic path length because of the law of large numbers (on the number of connected paths), and the distributions move strongly toward the lower bound [Lovejoy & Loch(2003)]. Also, by the

Central Limit Theorem the distribution of characteristic path lengths is well approximated normal as the random network size grows and/or the density grows. Under the condition that $N \times d \geq 25$, the rule of thumb is to set distribution mean as $2 - d$ and standard deviation $\sigma < 1$. In Figure 6 we plot the 95% probable range of clustering coefficients C for $d = 0.01$ and 0.02 in red and green set of lines respectively with the mean curve in the middle. Note that the standard deviations are much smaller for larger densities. Therefore, these 95% probable ranges are set apart far from each other in the case of clustering coefficients. Therefore, there is very little confusion identifying the equivalent density and size pair networks' clustering coefficients in the case of random graphs.

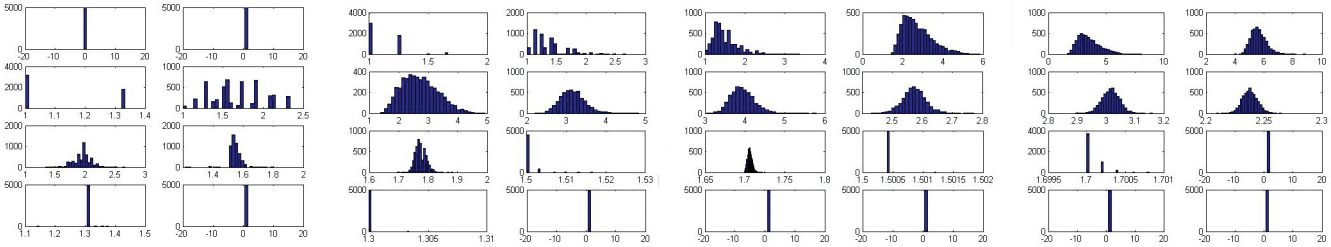


Figure 5: Distributions of characteristic path lengths by network size and density

For a smaller density networks $0 < d \leq 0.1$, we could get a consistent pattern of an inverse relationship with network size between 200 and 1,000 ($\frac{1}{N}$). The specific patterns of characteristic path lengths are described in following:

$$\begin{aligned}
&\text{when } p = 0.005, \quad L = 2.47 + 1982.96\frac{1}{N} && N = 500 - 1,000; \\
&\text{when } p = 0.01, \quad L = 2.37 + 847.29\frac{1}{N} && N = 200 - 1,000; \\
&\text{when } p = 0.02, \quad L = 2.30 + 323.38\frac{1}{N} && N = 100 - 1,000; \\
&\text{when } p = 0.05, \quad L = 2.13 + 97.62\frac{1}{N} && N = 25 - 1,000; \\
&\text{when } p = 0.10, \quad L = 1.87 + 32.97\frac{1}{N} && N = 25 - 1,000.
\end{aligned}$$

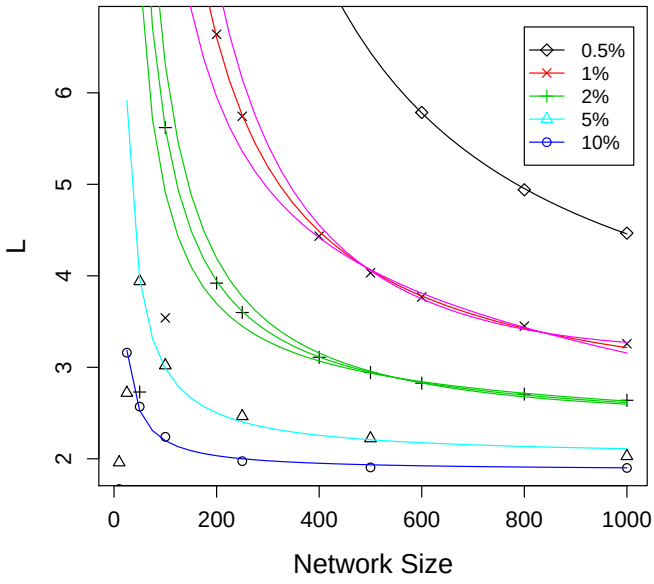


Figure 6: Iso-graph of characteristic path length means by density.

5.3 Interpretation of Network Centralizations

Centralizations, a graph-level measure for , range between a value 0 and 1. The larger the value is, the more likely it is that a single actor is quite central in the network with remaining nodes considerably less central. Hence, we may look at it as how heterogenous the network is or how variable each node's centralities are. Therefore we expect random graph's centralization values to be closer to 0 than 1 as all nodes behave more or less like the rest of the nodes.
([Newman et al.(2000)Newman, Strogatz, & Watts])

5.4 Betweenness Centralization

Betweenness centralization distributions are skewed leaning toward lower values, as we see in most of the following figures 9, 10, and 11. With a relatively constant 6:1 ratio of sample means to sample standard deviations, the distribution can be well approximated by a gamma distribution with the shape parameter $36 = 6^2$ and scale parameter that sets the mean as we calculated in Table 2 (see Appendix9.2 for details). From Figure 7, we find betweenness centralizations converging to a value regardless of the density (or the degree distribution) as the network size grows large. For example, when $N = 200$ and density 1%, the mean of 1250 random graphs' betweenness centralization is 0.0676, but the mean drops below 0.01 as the density grows to 5% and greater. On the other hand, when $N = 1,000$ and while $d > 1\%$, all graphs have means below 0.001. This is due to the nature of random graphs, where no particular nodes stand out as better connectors of any other two nodes in the network with a uniform probability of link formation. Hence, the

betweenness centralization occupies very small values with even smaller standard deviations. The largest mean predicted of betweenness centralization is 0.25 when $N = 25$ and $d = 0.1$ and the sample standard deviation 0.08.

The 95% probable range of betweenness centralization has an uneven length of wings about the expected mean. It should be about two standard deviations below the mean and 2.3 standard deviations above the mean due to its longer tail for the higher values according to the gamma distribution quantile profile. So we write it as $(\hat{\mu}_{(d,\ N)} - 2\hat{sigma}_{(d,\ N)}, \hat{\mu}_{(d,\ N)} + \frac{7}{3}\hat{\sigma}_{(d,\ N)})$. With the 6:1 ratio of $\hat{\mu}$ to $\hat{\sigma}$, we can simplify the expression as $\frac{2}{3}\hat{\mu}$, $\frac{11}{18}\hat{\mu}$, and we approximate $\hat{\mu}$ for $N \geq 200$ as:

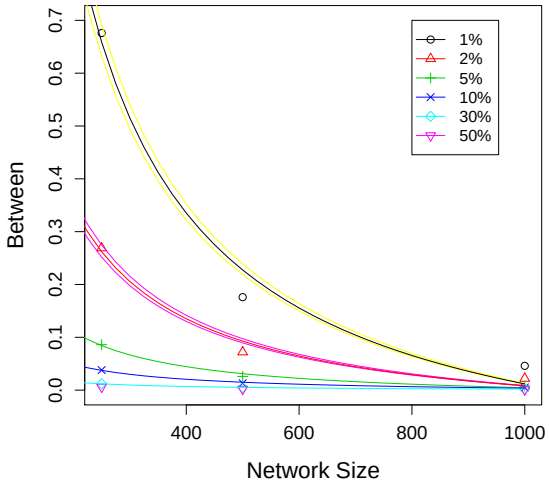
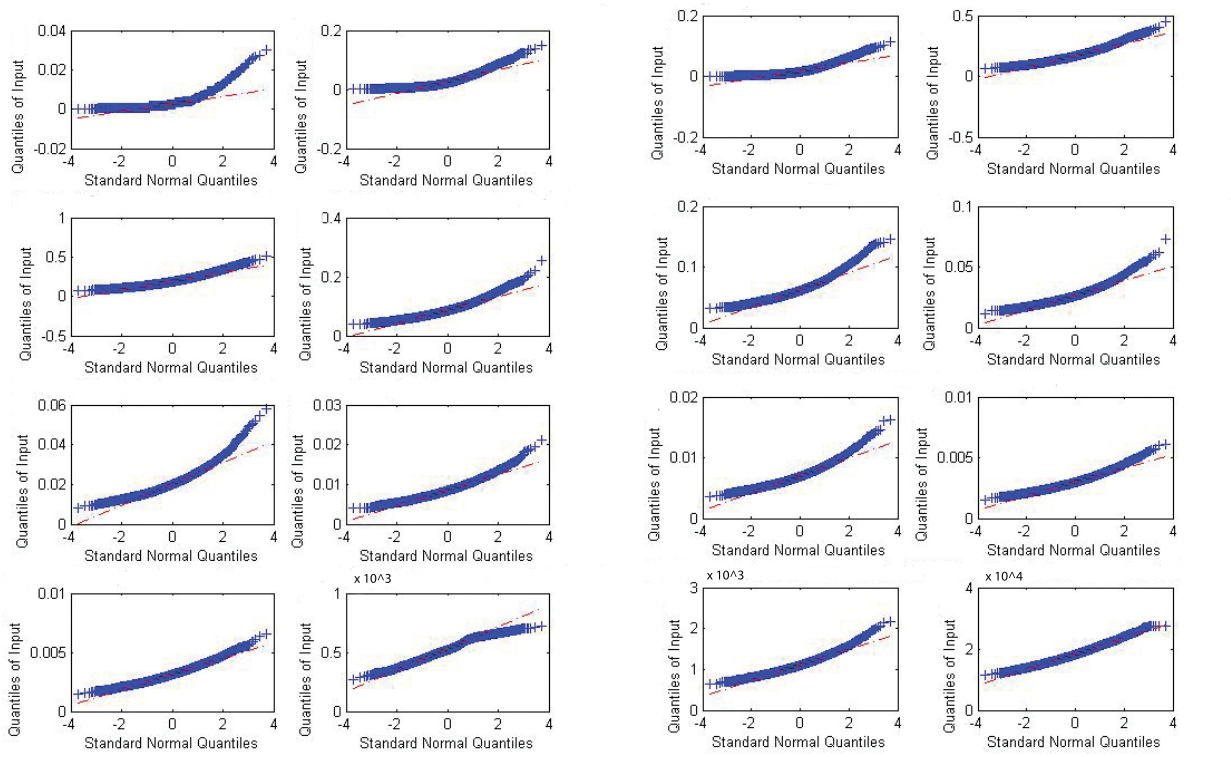


Figure 7: Betweenness centralization means against network size interpolated with an 1/N relationship

Table 2: Estimation of betweenness centralization means and standard deviations

Size N	Density d(%)	Mean $\hat{\mu}$	Std. dev $\hat{\sigma}$
200-1,000	1	$-0.20 + \frac{215.71}{N}$	$\frac{4.46}{N}$
	2	$-0.07 + \frac{84.64}{N}$	$\frac{1.72}{N}$
	5	$-0.02 + \frac{26.57}{N}$	$\frac{0.51}{N}$
	10	$\frac{11.21}{N}$	$\frac{0.21}{N}$
	30	$\frac{3.35}{N}$	$\frac{0.053}{N}$
	50	$\frac{1.71}{N}$	$\frac{0.027}{N}$



(a) N=50 (b) N=100

Figure 8: **Q-Q plot of betweenness centralization at different densities.** Each panel has the density growing from left to right and top to bottom. We see most of the Q-Q plots curved and deviating from normality as both tail ends lift away from the theoretical Standard Normal Quantiles.

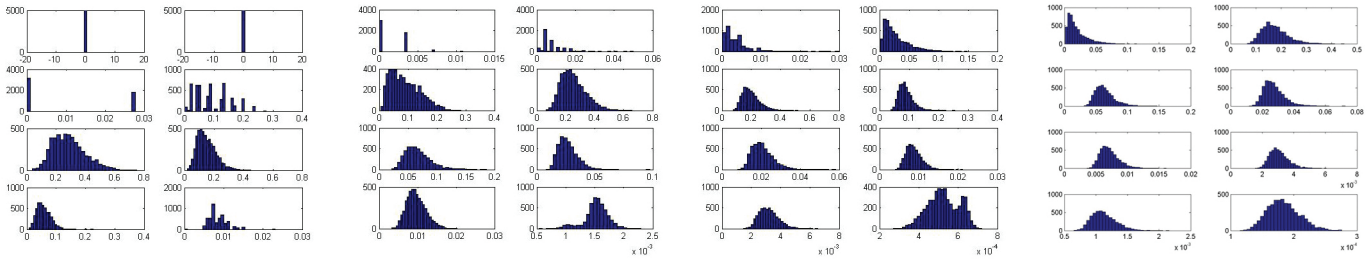


Figure 9: Histograms of betweenness centralization by network size and density where $N \leq 100$

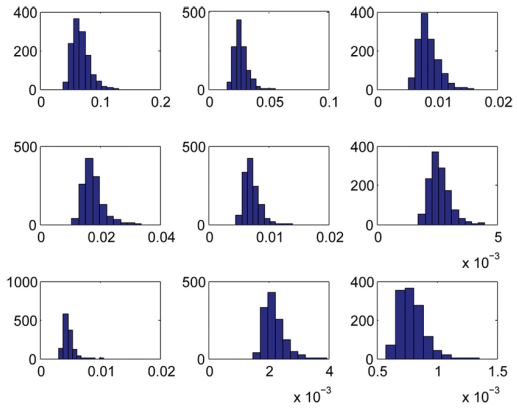


Figure 10: **Histograms of betweenness centralization by network size and density** First row is $N = 250$, second $N = 50$, and the last $N = 1000$, with first column $d = 1\%$, second 2%, and third 5%

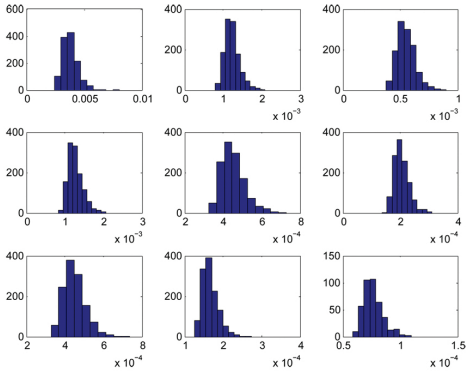
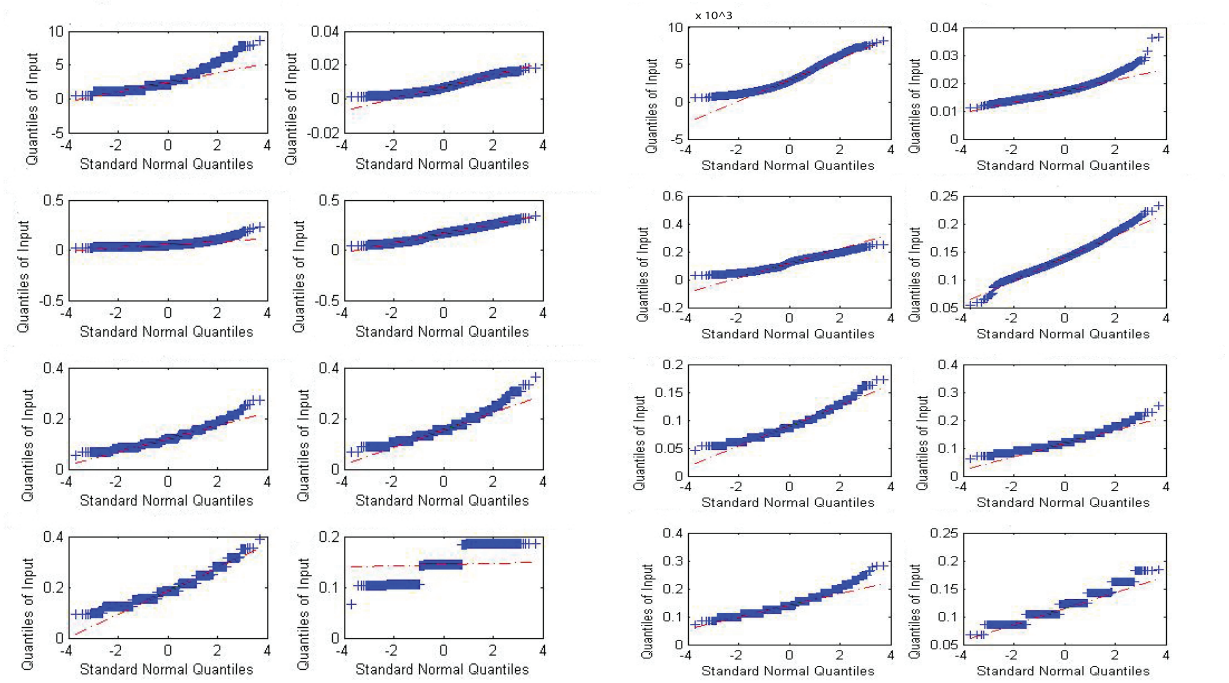


Figure 11: **Histograms of betweenness centralization by network size and density** First row is $N = 250$, second $N = 50$, and the last $N = 1000$, with first column $d = 10\%$, second 30%, and third 50%



(a) N=50 (b) N=100

Figure 12: **Q-Q plot of closeness centralization at different densities.** Each panel has the density growing from left to right and top to bottom. The distributions deviate from the theoretical Standard Normal Quantiles.

5.5 Closeness Centralization

For closeness centralization, the distributions are skewed leaning toward lower values (see Figures 13, 14, and 15) just like betweenness centralization distributions. But unlike the betweenness centralization, the distributions are multi-modal. We observe this from Q-Q plots in Figure 12a and 12b, where the sample values appear like a step-function. A reason for observing such stacks of points is due to the small deviations from the expected connectedness of the networks. When the density gets greater, it translates to a higher probability of link existing between two nodes. Hence, the expected number of connected nodes have little deviations, which results into having repeated values.

For the 95% probable range of closeness centralization, we may set out an even length of wings about the expected mean because the Q-Q plots show that the deviation from the standard normal quantiles start around or beyond the $\pm 2\sigma$ (standard deviations) reading it from the x-axis. To be more specific, the whole range spans two standard deviations about the mean, and we write it as $(\hat{\mu}_{(d,\ N)} \pm 2\hat{sigma}_{(d,\ N)})$. With little clues to take the expectation of the closeness centralizations, we just use the spline methods to interpolate the values for different densities shown in Figure 13.

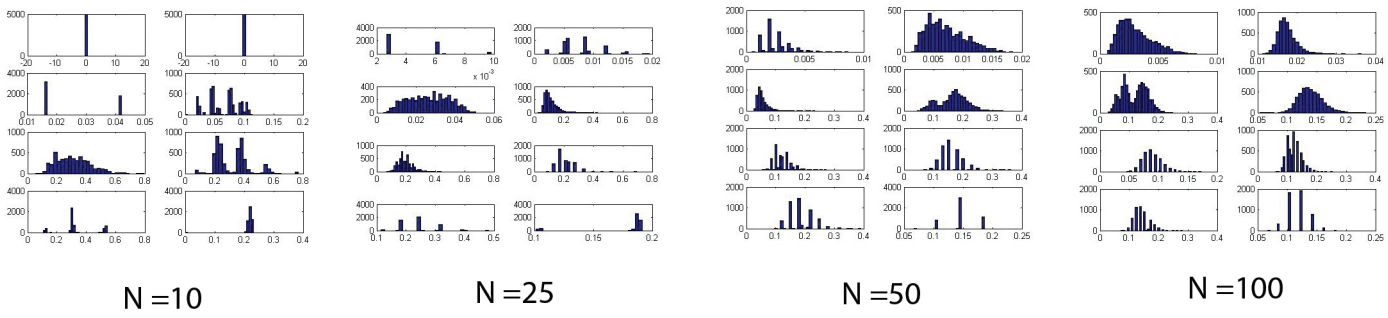


Figure 13: **Histograms of closeness centralization by network size and density**jk

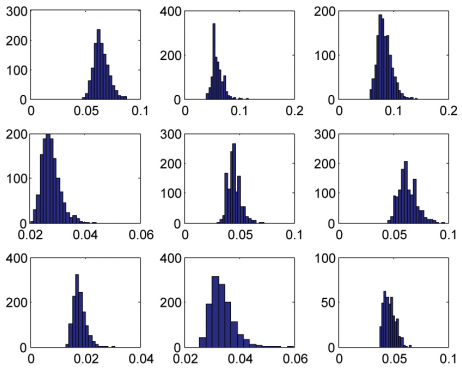


Figure 14: **Histograms of closeness centralization by network size and density** First row is $N = 250$, second $N = 50$, and the last $N = 1000$, with first column $d = 1\%$, second 2%, and third 5%

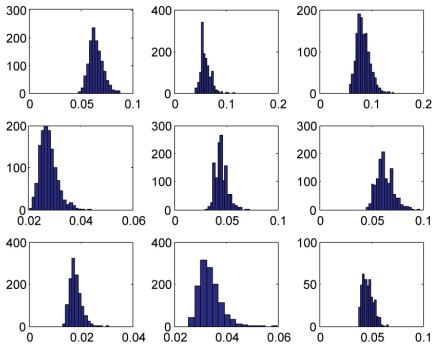


Figure 15: **Histograms of closeness centralization by network size and density** First row is $N = 250$, second $N = 50$, and the last $N = 1000$, with first column $d = 10\%$, second 30%, and third 50%

The 95% probable ranges have little overlap when $d \geq 0.1$ as the network size differs. We should further investigate the reasons behind this.

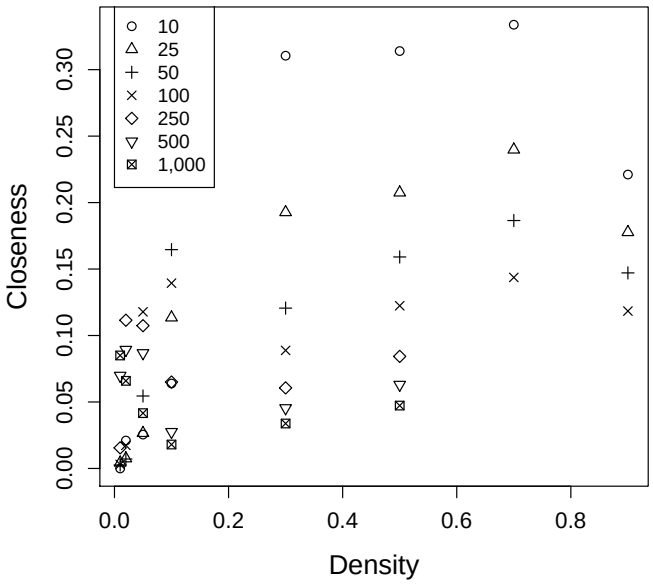


Figure 16: **Closeness centralization mean values by network size and density** The greater the network size, the smaller the closeness centralization values are. mean dips down when the network

6 Probable Ranges of Undirected Network Node-level Metrics

In this section we explain the derivation of distributional quantities of a variety of network metrics on node level, for random graphs node size 25-1,000. These graphs are assumed to be random, which means that all nodes have an equal probability of being connected to all other nodes, therefore all nodes in a network behave similarly as one another.

As we increase the network size, the overall shape of the metric distributions are not only smoother on a continuous scale but also more consistent (with a tighter range of metric values, hence smaller standard deviations). We capture such a tendency in the following analysis assuming a fixed density for a set of networks and increasing the network size. However, degree centrality and centralization are exceptions in fixed density because in a random graph generation there is a clear relationship between degree centrality/centralization and the network density and size pair.

In all graphs: 'Eig' is for eigenvector centrality; 'Bet' is betweenness centrality; and 'Clo' is closeness centrality. These graphs show the generated networks' centrality means and the standard deviations with overall trends shown in blue and green curves. We run a regression on every pair of network centrality measures and density we have by randomly selecting 250 values from each network size data. The plotted values are the overall mean of network centrality measures with a fixed density and size. For standard deviations, we run a regression on four up to seven values that we obtained from the centrality values of $250 \times N$ data points. (See more in **Appendix-More Graphs** section.)

6.1 Degree Centrality

The number of links connected to a node is called the degree k of that node, and has a probability distribution p_k given by

$$p_k = \binom{N}{k} p^k (1-p)^{N-k} \simeq \frac{d^k e^{-d}}{k!}$$

where the second equality becomes exact in the limit of large N. This distribution we recognize as the Poisson distribution. Hence, the random graph's degree centrality follows a Poisson distribution.
When we normalize the degree centrality by dividing node degrees by the graph size, N, we rather obtain a continuous distribution of the normalized degree centrality as $N \to \infty$, instead of a discrete distribution. By the Central Limit Theorem we also approximate the normalized degree centrality distribution $\frac{1}{N} Bin\ (N, p)$ by $Norm\ (p, \frac{p(1-p)}{N})$.

For a symmetric adjacency matrix, the in-degree and out-degree are the same, as well as the row and column degree and the total degree centrality. So, we take a look at the total degree centrality to represent all the degree distributions.
As it is derived above, the mean of the normalized degree centrality is the network density p, and the variance is $\frac{p(1-p)}{N}$ where N is the network size. For the 95% most probable interval of random

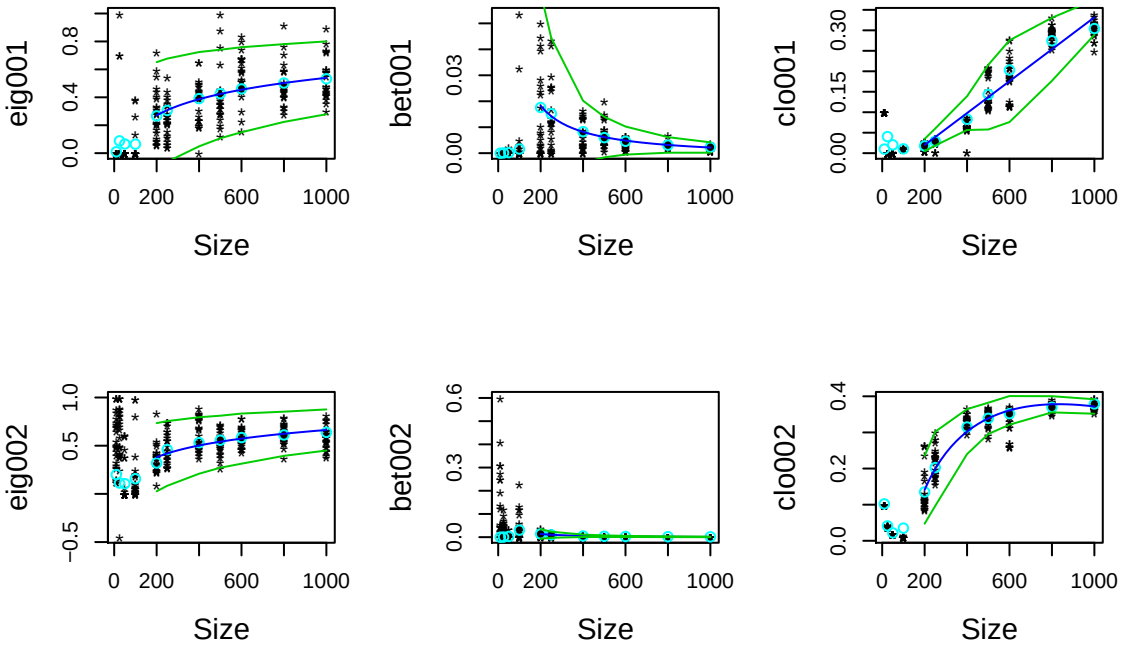


Figure 17: **Standard network metrics' 95% probable intervals**: (expected mean $\pm$ 2(standard deviation)) for random networks of size 200-1,000 at density 1% (top) and 2% (bottom row).

graph node's mean behavior, we get $p \pm 2\sqrt{\frac{p(1-p)}{N}}$. In other words, I can be sure that the 95% of the random graph nodes' degree centrality will fall in between $p - 2\sqrt{\frac{p(1-p)}{N}}$ and $p + 2\sqrt{\frac{p(1-p)}{N}}$. For example, if we have $N = 100$ and $p = 0.05$, then the 95% probable interval for a node's degree distribution is (0.0064, 0.0936).

6.2 Eigenvector Centrality

Another way to measure the importance of a node in a network is to obtain a positive eigenvector for adjacency matrix and collect the eigenvalues. These values ranging between -1 and 1 show relative scores of nodes where high-scoring nodes have links that are central connectors to the rest of the network than low-scoring nodes where the links may be many but isolated ones.

Table 3: Estimation of eigenvector centrality mean and standard deviation

Density	size (d, N)	Mean	Standard deviation
0.005	500-1,000	$0.1549 + 2.546 \times 10^{-4}\ N$	$0.1874 - 3.656 \times 10^{-5}\ N$
0.01	200-1,000	$-0.6147 + 0.1671\ \ln(N)$	$0.1724 + \frac{5.787}{N} 4.898 \times 10^{-5} N$
0.02		$-0.5499 + 0.17559\ \ln(N)$	$0.1390 + \frac{9.518}{N} 4.312 \times 10^{-5}\ N$
0.05	10-1,000	$-0.1916 + 0.1367\ \ln(N)$	$0.5375 - 0.0699\ \ln(N)$
0.1	10 -1,000	$0.1208\ \ln(N)$	0.30
	(200-1,000)		$0.0443 + 14.0071\ \frac{1}{N}$
0.3		$0.3380 + 0.0829\ \ln(N)$	0.31
			$0.0224 + 8.3643\ \frac{1}{N}$
0.5		$0.4972 + 0.0659\ \ln(N)$	0.28
			$0.0133 + 5.3929\ \frac{1}{N}$
0.7	10-100	$0.7843 + 5.615 \times 10^{-4}\ N$	
0.9		$0.9049 + 2.579 \times 10^{-4}\ N$	

The overall distribution is not smooth nor normal, but the middle 75% behaves very much like a normal where we observe this from Q-Q plots (even from the $N \leq 100$). See Figure x. Hence, we give a smaller 65% probable range for this centrality measure by providing $(\hat{\mu}_{(d,\ N)} \pm \hat{sigma}_{(d,\ N)})$. We already encounter trouble distinguishing which density range the potential eigenvector centrality measure mean should fall into as the standard deviation is around 0.1, if not greater for $200 \leq N \leq 1,000$. See Figure 18 where the iso-line intervals are only about 0.1 apart.

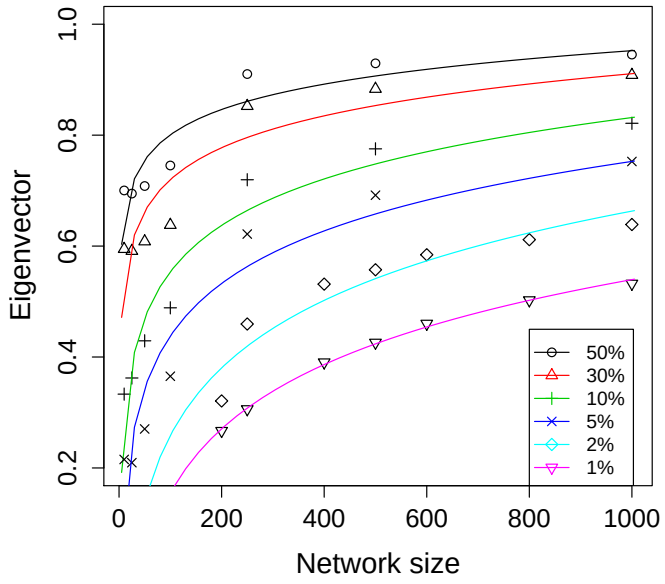


Figure 18: Iso-graph of eigenvector centrality means by density.

6.3 Betweenness Centrality

Betweenness centrality distributions are well approximated with a normal distribution when $N \geq 25$ and $d \geq 0.4$. However, note that when $N = 25$ and $d = 0.5$, the left tail is cut off by 0.

$$\mu_{bet}(d, N) = \frac{1-d}{N}.$$

$$\begin{aligned}
\sigma_{bet}(d, N) &= 0.0019 * 3^{\log_2^{\frac{100}{N}}} && \text{when } d = 0.4 \\
&= 0.0011 * 3^{\log_2^{\frac{100}{N}}} && \text{when } d = 0.5 \\
&= 0.0004 * 3^{\log_2^{\frac{100}{N}}} && \text{when } d = 0.7 \\
&= 0.0001 * 3^{\log_2^{\frac{100}{N}}} && \text{when } d = 0.9
\end{aligned}$$

When d is small (that is, $d \leq 0.1$), there are too many isolated nodes in networks. Hence, the betweenness centrality piles up near 0. Only when $N \times d \geq 25$, the distribution starts spreading out evenly to the left and right.

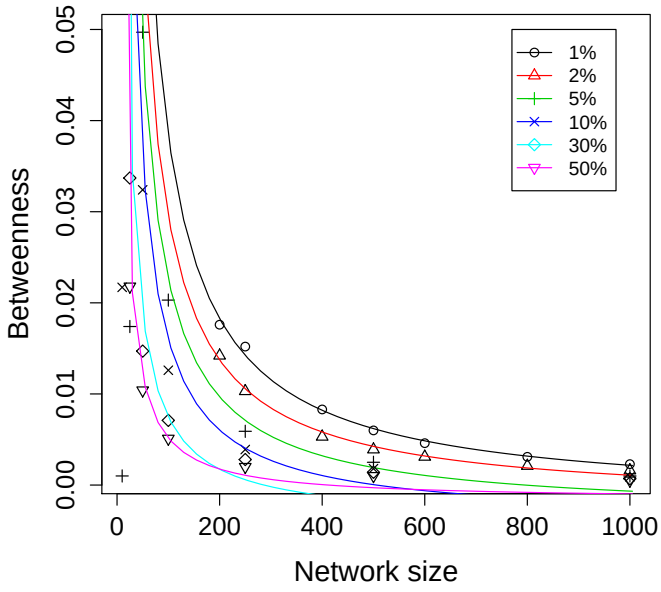


Figure 19: Iso-graph of betweenness centrality means by density.

6.4 Closeness Centrality

Normality is observed, when $d = 0.3, 0.5, 0.7,$ and 0.9 (i.e. d is greater than 0.3)

$$\begin{aligned}
\mu_{clo} &= (\text{characteristic path length})^{-1} \\
\sigma_{clo} &= \frac{1.60}{N} \quad \text{when } d = 0.3 \\
&= \frac{2.22}{N} \quad \text{when } d = 0.5 \\
&= \frac{2.70}{N} \quad \text{when } d = 0.7 \\
&= \frac{2.46}{N} \quad \text{when } d = 0.9
\end{aligned}$$

When d is small (that is, $d < 0.1$), we observe many isolated nodes in graphs. Hence, the closeness centrality of nodes from a network size N is heavily piled on $\frac{1}{N}$ and $\frac{1}{N+1}$. We find more interesting distributions, when $N = 25$ and $d = 0.1$, $N = 50$ and $d = 0.05$, $N = 100$ and $d = 0.02$, there are not only huge spikes at $\frac{1}{N}$ but also another peak to the right of $\frac{1}{N}$.

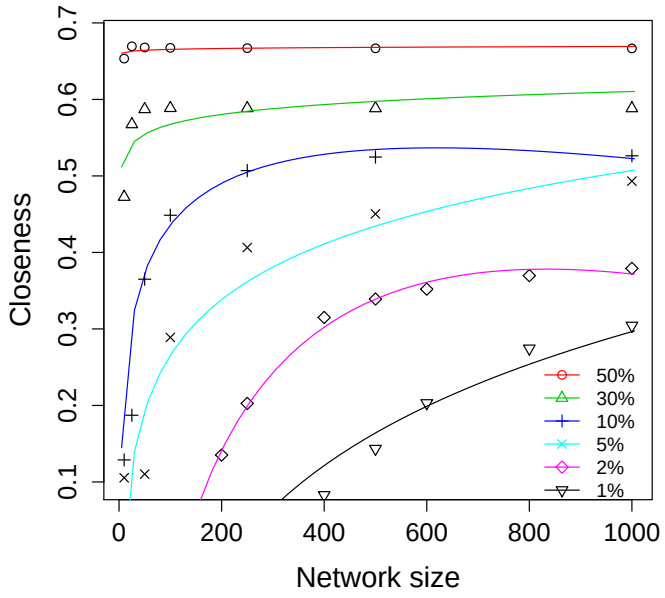


Figure 20: Iso-graph of closeness centrality means by density.

7 Non-square Adjacency Matrix Network Metrics' Distributions

We encounter a case where personnel is linked to resource or knowledge, each representing row or column. The number of rows and the number of columns may differ. So, we represent this case in a K x N adjacency matrix where each entry (i, j) has 1 for the presence of a link and 0 for the absence of a link between i^{th} and j^{th} nodes. Such adjacency matrices are obviously not symmetric, that is a 1 in (i, j) entry does not imply a 1 in (j, i) entry and vice versa.

7.1 Degree Centrality and Centralization

We would like to examine whether links in real scenarios occur more or less at random or not. Hence, we derive the distributions of degree metrics of non-square random networks where we assume a presence of a link at random across the matrix cells (i.e. the presence of a link is equiprobable for all pairs of nodes). Then we summarize its row degree and column degree distributions and provide the 95% most probable interval of each metric.

The derivation is similar to the symmetric adjacency matrix case. We fix a node from the row lineup and know that its probability distribution of the degree (number of links or neighbors) follows a Binomial distribution with a fixed possible number N linkages and probability p. Hence, the row degree centrality would follow a $Bin(N, p)$. If we normalize row degree centrality, the

distribution will be divided by N, and the expected mean centrality is p, and the standard deviation is $\sqrt{p(1-p)}$. By the same reasoning, the column degree centrality would follow a $Bin(K, p)$ because a column node could link to each K elements with a fixed probability p. If we normalize column degree centrality, the distribution will be divided by K, and the distribution is the same as that of the normalized row degree centrality. See Table4 for summary.

Table 4: K x N Non-square matrices normalized degree distribution summary

Degree metrics	Distribution	Mean	Standard Deviation
Row degree centrality	$\frac{1}{N}Bin\ (N, p)$	p	$\sqrt{p(1-p)}$
Row degree centralization	$Norm\ (Np, \frac{Np(1-p)}{K})$	p	$\sqrt{\frac{Np(1-p)}{K}}$
Column degree centrality	$\frac{1}{K}Bin\ (K, p)$	p	$\sqrt{p(1-p)}$
Column degree centralization	$Norm\ (Kp, \frac{Kp(1-p)}{N})$	p	$\sqrt{\frac{Kp(1-p)}{N}}$

8 Final Remarks

9 Appendix

9.1 Notation

- N: network size.
- d: network density.
- p: probability $0 \leq p \leq 1$.
- Norm(μ, σ^2): normal distribution with mean μ and standard deviation σ.
- *Eig/ eig*: Eigenvector (centrality);*Bet/ bet*: Betweenness (centrality); *Clo/ clo*: Closeness (centrality).

9.2 Terminology

- The terminology **graph** and **network** are used interchangeably. In most cases we tried to stick to use the term *network so that there is less confusion in our daily use of language. However, in the literature of mathematics, random graph is coined by mathematicians Erdos*

and Renyi and has been widely used ever since. Therefore, occasionally graph will mean a general network.

- *A **digraph** is a directed graph.*

- *Gamma($\alpha, \frac{1}{\beta}$) probability density function:*

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp^{-\beta x}, \quad 0 \le x < \infty, \;\; \alpha, \beta > 0$$

$$E[X] = \frac{\alpha}{\beta}, \;\; Var[X] = \frac{\alpha}{\beta^2}$$

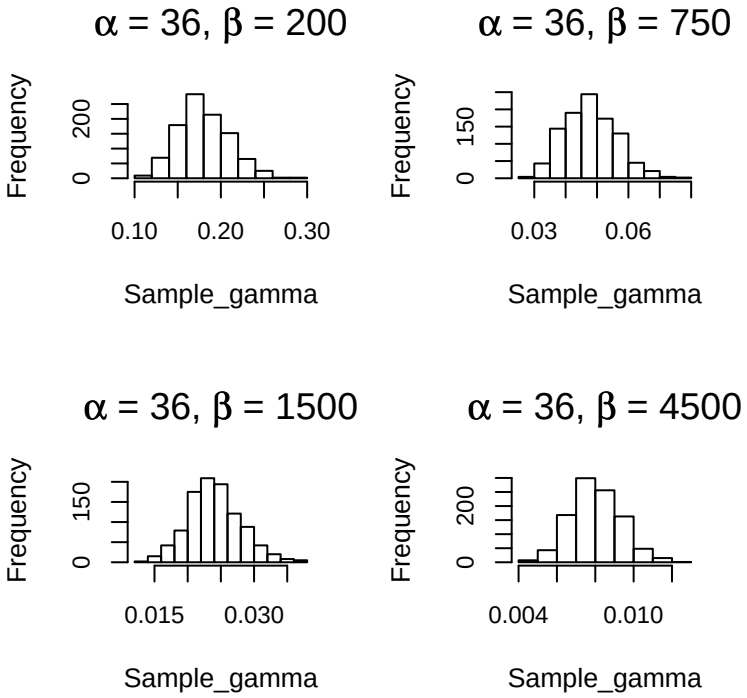


Figure 21: **Gamma distributions with various parameter values α and β.** These are skewed toward lower values.

If we draw 1,000 sample from a gamma distribution with the shape parameter $\alpha = 36$ and the scale parameter β such that $\frac{\alpha}{\beta}$ is the expected mean, we may set β to ...

9.3 Additional Graphs

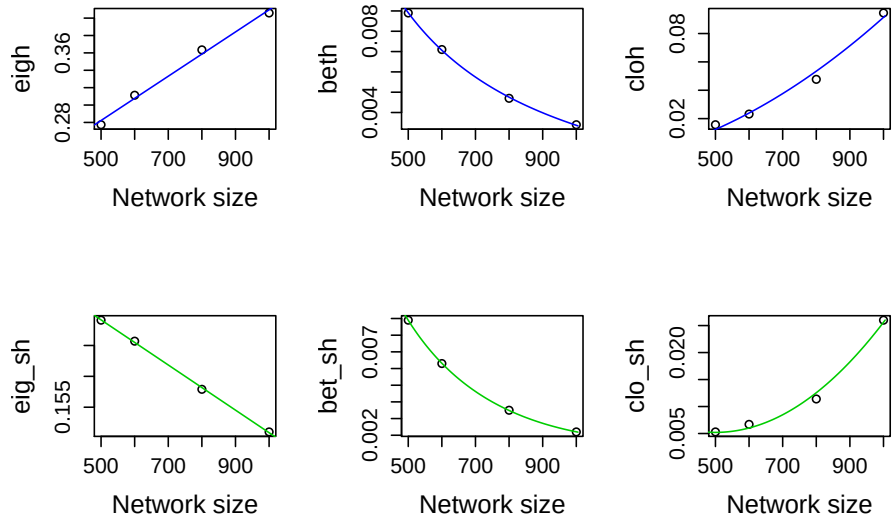


Figure 22: **Network measure means and standard deviations against network size when d=0.005.** The first row shows when the measures' means, and the second row is the variances.

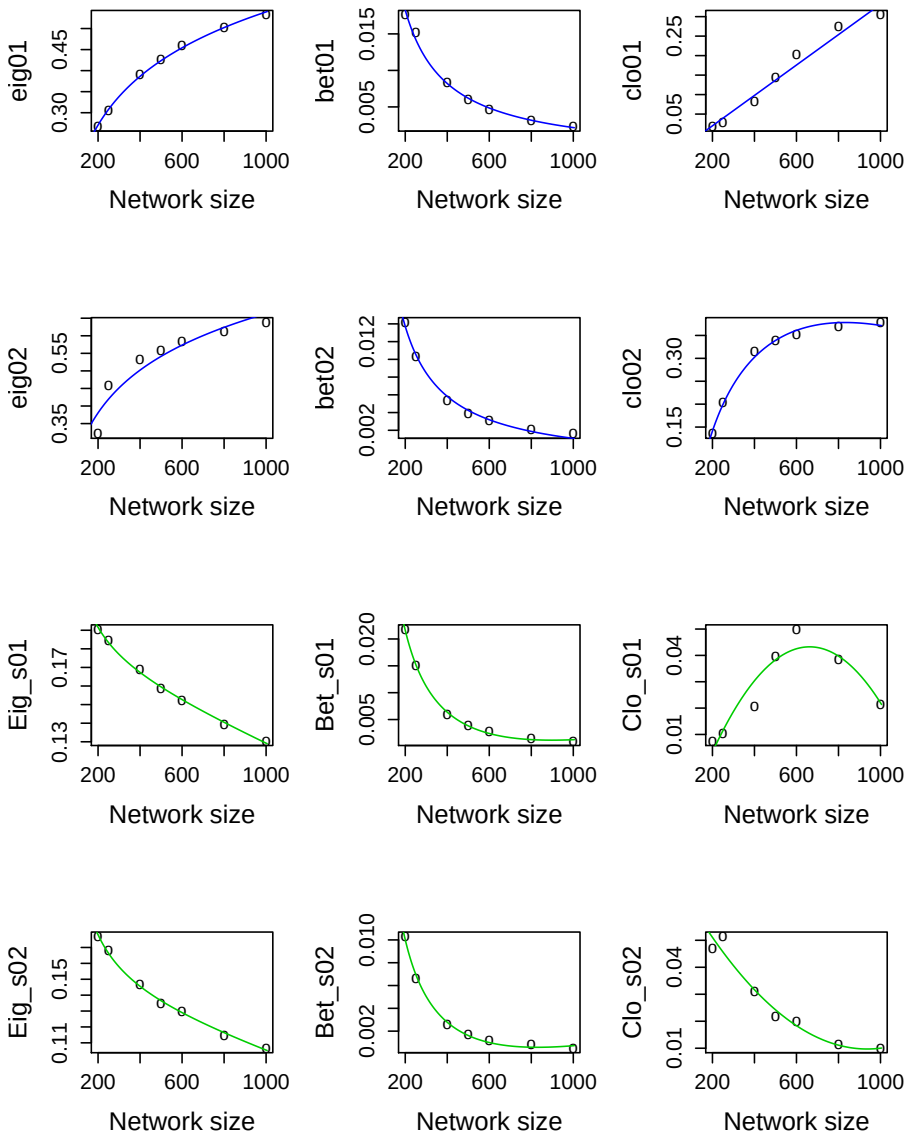


Figure 23: **Network measure means and standard deviations against network size when d = 0.01 and 0.02.** The top half shows the measures' means, and the bottom half reports the standard deviations. The first and the third rows are when the density is fixed at 1%, and the second and the fourth rows are at 2%. For the mean interpolations: eigenvector centrality is well explained by a log transformation of network size; betweenness centrality is very well captured by 1/(Network size) transformation; closeness centrality holds a linear relationship at first then a logarithmic transformation of network size.

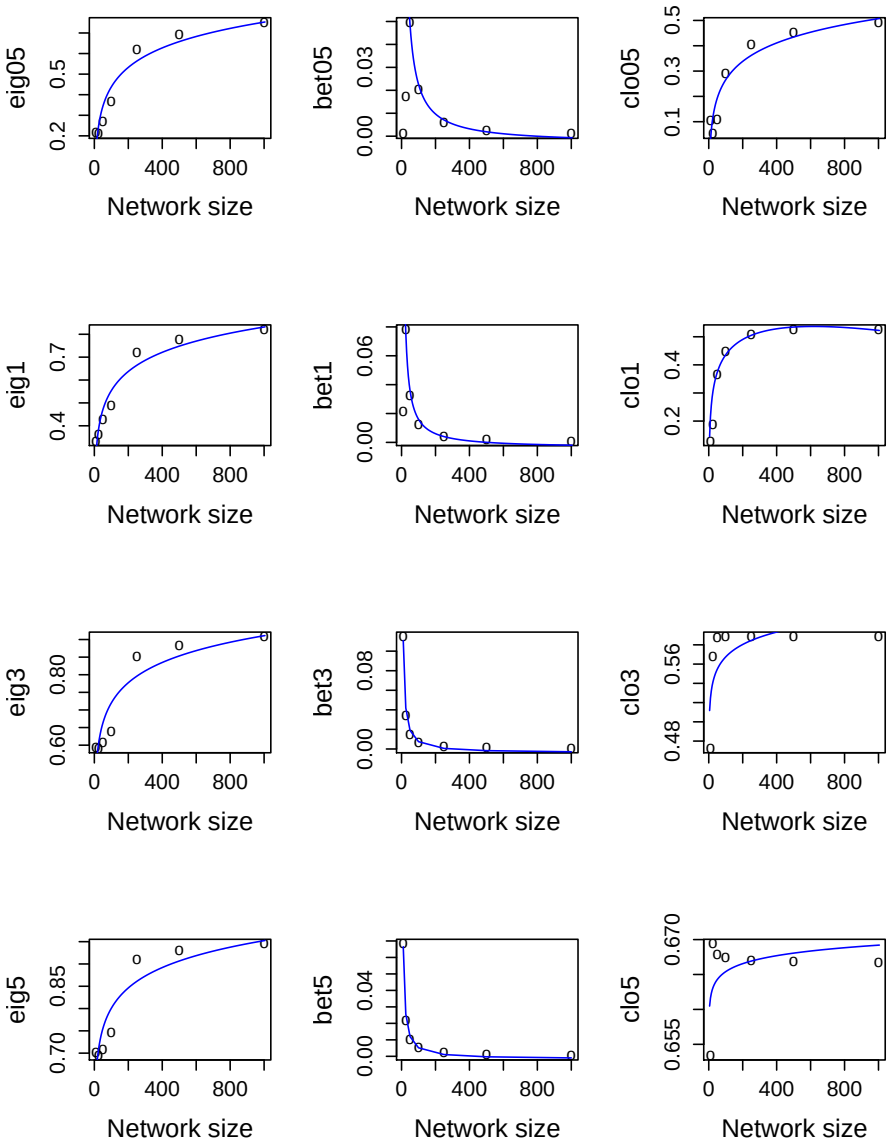


Figure 24: **Network measure means against network size when d = 0.05, 0.1, 0.3, and 0.5.** The first row shows when the density is fixed at 0.5%, the second row is at 10%, the third at 30%, and the last row is at 50%. Eigenvector centrality has values plateauing as the network size grows, yet a log transformation is not a great fit; betweenness centrality is very well captured by 1/(Network size) transformation except when the density is low and network size small; closeness centrality seems to also plateau at a certain point, but the relationship is yet uncertain.

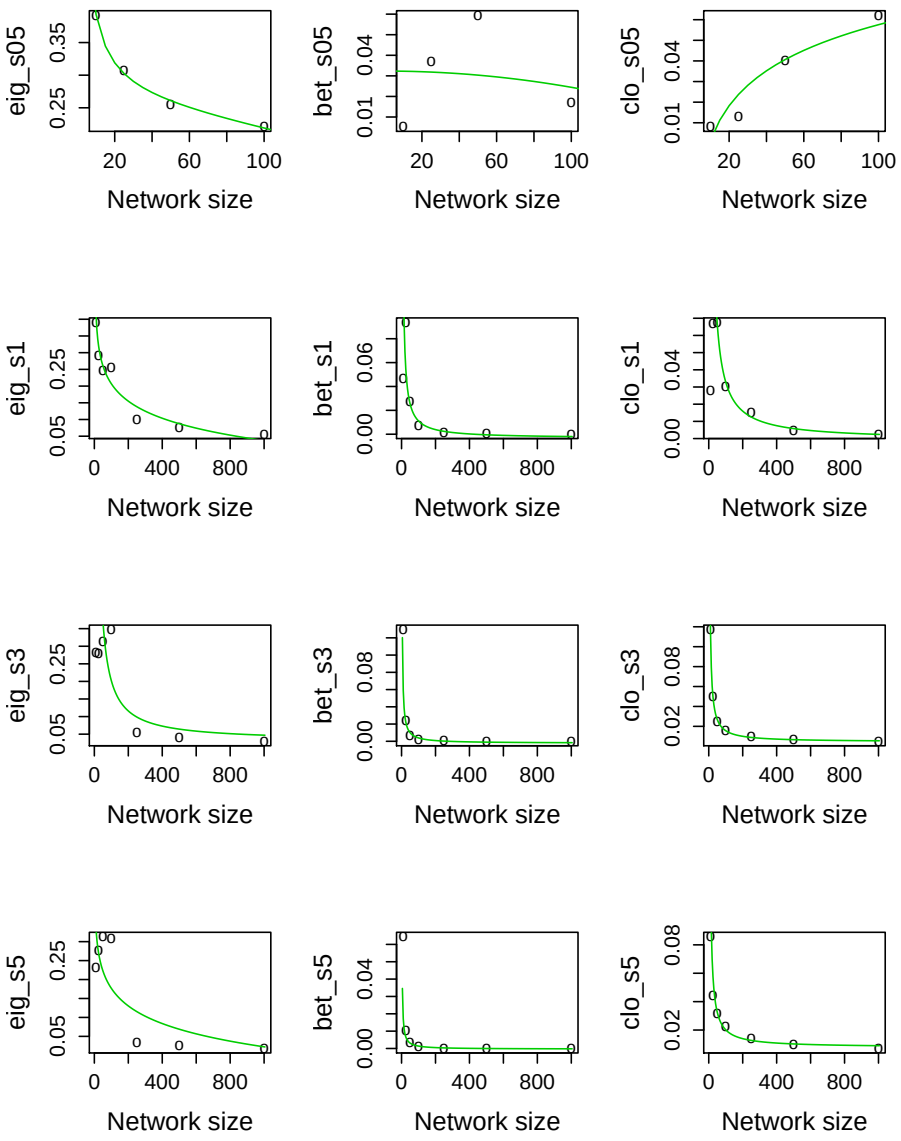


Figure 25: **Network measure standard deviations against network size when d = 0.05, 0.1, 0.3, and 0.5.** The first row shows when the density is fixed at 5%, the second row is at 10%, the third at 30%, and the last row is at 50%. Eigenvector centrality stdev. decreases as 1/(Network size) with increasing size; betweenness centrality and closeness centrality stdev. also show a similar behavior except when the network density is at 5%; when the network size is small (¡100) the standard deviation grows on log(N) scale as we see in the top right.

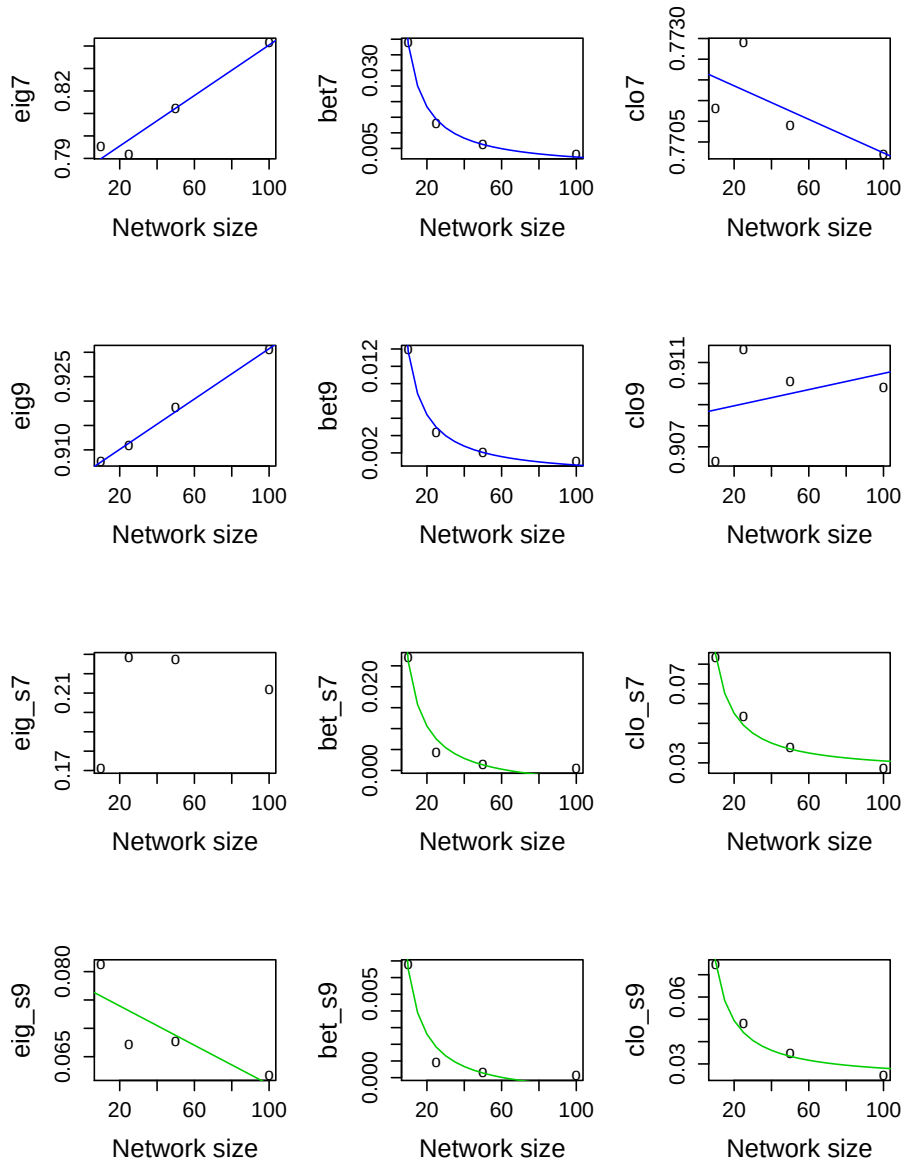


Figure 26: **Network measure means against network size when d = 0.7 and 0.9** The top half shows the measures' means, and the bottom half reports the standard deviations. The first and the third rows show when the density is fixed at 70%, and the second and the fourth rows are at 90%. Since the density is very high, we only look at network size up to 100. For this small range of sizes, means of eigenvector centrality grows linearly with network size; that of betweenness centrality is very well captured by 1/(Network size); however, closeness centrality does not show a clear relationship. For standard deviations, eigenvector centrality gives no clear relationship ; betweenness and closeness centralities are explained by 1/(Network size) overall.

References

[Lovejoy & Loch(2003)] W. Lovejoy & C. Loch (2003). 'Minimal and maximal characteristic path lengths in connected sociomatrices.' *Social Networks* **25**:333–347.

[Newman et al.(2000)Newman, Strogatz, & Watts] M. Newman, S. Strogatz, & D. Watts (2000). 'Random graphs with arbitrary degree distribution and their applications.'

[Wasserman & Faust(1994)] S. Wasserman & K. Faust (1994). *Social network analysis*. Cambridge University Press, Cambridge.