**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

| | |
|---|---|
| Cyra Christina Mehta | Date |

Centralization in Small Graphs

By

Cyra Christina Mehta

Doctor of Philosophy

Biostatistics

---

Vicki Stover Hertzberg, Ph.D.
Adviser

---

Michael Haber, Ph.D.
Committee Member

---

Andrew Hill, Ph.D.
Committee Member

---

Lance A. Waller, Ph.D.
Committee Member

---

Howie Weiss, Ph.D.
Committee Member

Accepted:

---

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

---

Date

Centralization in Small Graphs

By

Cyra Christina Mehta
M.S., Emory University, 2013
M.S.P.H., Emory University 2004
B.A., Emory University, 2000

Adviser: Vicki Stover Hertzberg, Ph.D.

An Abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2014

Abstract

Centralization in Small Graphs

By Cyra Christina Mehta

Network science provides a new and valuable set of techniques for investigating problems by providing a different perspective. One important network characteristic is centralization, which is a quantification of how much one node dominates the network according to a particular measure of node influence. Centralization is a graph-level measure and can be calculated for closeness, betweenness, and degree. This research investigated the range of centralization values obtained by common graph generating methods, properties of centralization, and whether centralization is associated with disease spread.

The first study examined the centralization values and graph structures produced by Erdös-Rènyi $Gnm$ random graphs, Barabási-Albert preferential attachment graphs, small world graphs, and two node preferential attachment graphs as well as Star Start, a new graph generating method that produces graphs across the full theoretical range of centralization values. Erdös-Rényi random graphs produce low to moderately centralized graphs and small world graphs are only low to moderately centralized. Barabási-Albert preferential attachment graphs can be highly centralized but do not produce a variety of graph structures. Two node type preferential attachment and Star Start produce most of the full range of centralization values with a broad range of structures.

Using the Star Start program, the second study explored the properties of centralization, including prediction based on average or maximum node centrality. Correlation between centralization measures decreases as graph order increases. Models predicting centralization based on maximum centrality perform reasonably well, especially when the maximum centrality value $\leq 0.6$. Models based on average centrality fit poorly after the average increases past the average centrality of a star graph.

Lastly, the association between centralization and epidemiologic endpoints using a Susceptible-Infected-Recovered ($SIR$) compartment model of disease spread was examined. As degree or betweenness centralization increases the peak number of infected nodes increases, time until the peak decreases, and the final cumulative number of infected nodes also increases. Closeness centralization does not have as strong of a relationship and should only be considered for connected networks. The results also confirm that infecting the most central node first produces a more severe epidemic than randomly selecting a node.

# Centralization in Small Graphs

By

Cyra Christina Mehta
M.S., Emory University, 2013
M.S.P.H, Emory University, 2004
B.S., Emory University, 2000

Adviser: Vicki Stover Hertzberg, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2014

# Acknowledgments

I would like to gratefully acknowledge all of the support and encouragement I have received from my family, friends, and classmates. My dissertation adviser Dr. Vicki Hertzberg provided invaluable guidance during this process and has also served as an excellent mentor. I would also like to thank my committee members Dr. Lance Waller, Dr. Michael Haber, Dr. Andrew Hill, and Dr. Howard Weiss for their thoughtful comments that led to improvements in my research.

# Contents

# List of Figures

# List of Tables

# List of Movies

# Part I

# Background

# Chapter 1

# Introduction

## 1.1 Introduction

There are many definitions of a network, but a simple one is that a network is any group of items or people that can be connected by a common attribute or activity. A more technical definition is "a collection of points joined together in pairs by lines."[1] Networks include everything from the internet to the electric power grid to communication networks. Networks can be classified into four broad categories: technological (such as the internet, transportation/highway, or electric power grid), social (such as friendship or sexual partner), information (such as scientific citation or email), and biological (such as neural or metabolic).[1] This paper will utilize the framework of a communication network between people for ease of description, but the ideas can be generalized to any network.

It is accepted that the behavior of individuals cannot be easily understood without also considering the context of their surroundings and personal attributes. According to Valente, "Relationships influence a person's behavior above and beyond the influence of his or her individual attributes."[2] Indeed, traditional analytic methods include ways to investigate and control for the influence of covariates on the outcome. A critical assumption in traditional statistical analysis is independence between subjects. However, in networks the structure of the network, meaning the relationships between people, is important and must

be considered. The field of network science tries to answer questions about the behavior of one individual embedded in a group of others based on their location in the network. In other words, "Network analysis, thus, provides public health with a new way of framing and answering important health questions."[3] The methods of network science have been and are currently still being developed to adequately incorporate the structure of a network into an analysis and many questions need to be explored.

Given how useful network science can be in the investigation of the relationships between individuals, it is not surprising that network analysis techniques have been used by public health researchers. A recent review article suggests that researchers interested in three categories of public health networks use network analysis: transmission networks (either disease or information), social networks, and organizational networks.[3]

## 1.2 Definitions

For this paper a particular network will be called a *graph*, *nodes* are the points in the graph and *edges* connect the nodes. The number of edges in a graph, $m$, is called the graph's *size* while the number of nodes in a graph, $n$, is called the graph's *order*. The set of all nodes of a graph $g$ is $V(g)$ while the set of all edges is $E(g)$. A *path* is the sequence of edges between two nodes. The *geodesic* is the shortest path between two nodes, where distance is defined as the count of number of edges. Note that a geodesic need not be unique so there may be multiple geodesics between two nodes and these geodesics may or may not contain some of the same nodes. In the communication network framework, nodes are individuals and edges are communication lines. See Figure 1.1 for an example graph, a star graph of order 5.

A graph is *connected* if there is a path between each node in the graph and *disconnected* otherwise. A *component* is an isolated subgraph which has a path between all nodes in the subgraph. For example, a graph with one disconnected node has two components, a one-node component and another component where the remaining nodes are connected. A *community* or *cluster* is a densely connected group of nodes within a large, sparse network.

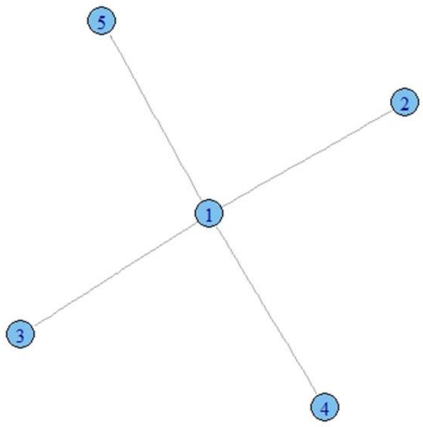


Figure 1.1: An example graph, a star graph with 5 nodes

A *clique* is the largest group of nodes in a graph that are fully connected with each other. A related idea is *k-core* which is the largest group of nodes that are connected to at least $k$ others.

A *digraph* is a directed graph, where the edges are directed from one node to another. Otherwise, edges are *undirected* or bi-directional. *Self-loops* are edges that start and end with the same node. Graphs with *multiedges* have more than one edge between nodes and are called *multigraphs*. Edges can have *weights* assigned to reflect their value or strength. Graph *density* is the number of edges in the graph over the total number of edges that are possible. A graph is *simple* if there is only one edge connecting nodes (multiple or weighted edges and self-loops are not allowed).

An *adjacency matrix*, $A$, of a simple graph with $n$ nodes, where the nodes are labeled from 1 to $n$, is an $n$ by $n$ matrix that contains a 1 in cell $A_{ij}$ is there is an edge between nodes $n_i$ and $n_j$ and a zero otherwise. In an undirected graph the adjacency matrix is symmetric and can be asymmetric for a directed graph. The *spectral radius*, $\rho$, of a graph is the largest nontrivial eigenvalue of the adjacency matrix.[4] The adjacency matrix can be modified to include self-loops, weighted or multiedge graphs, and directed graphs. The Laplacian of a graph is the matrix $L = D - A$ where $D$ is the diagonal matrix of node degrees and $A$ is the adjacency matrix.

In a *static* graph the number of nodes and contacts between nodes do not change while a

*dynamic* graph has properties that change over time, such as the number of nodes or edges or the relationship between nodes. A *bipartite* or *affiliation* graph, has nodes in two distinct groups and edges link the two groups. An example of a bipartite graph is club membership, with one set of nodes comprising the individuals and the other set of nodes comprising the possible clubs and edges linking individuals to the clubs they are members of.

*Regular* graphs are comprised of nodes with the same degree. Examples of a regular graph are lattice, ring, and complete graphs. Graph *isomorphisms* refer to graphs that have the same structure although the edges creating the graph are different. An example of this is a star graph with $n$ nodes. Each of the $n$ nodes can be the center of the star graph so there are $n$ isomorphisms.

In many situations, the number and strength of relationships between people is unrelated to their physical proximity to each other. Network analysis is ideally suited to allow an examination of these relationships because networks do not necessarily represent the actual physical location and distance between nodes. However, in spatial networks the nodes of the graph are located in space and distance between nodes influences possible connections between nodes. In these networks, the probability of an edge connecting two nodes is based on the Euclidean distance between the nodes. This representation is not typical and will not be discussed further.

## 1.3 Measures

Network science is distinguished by the attention given to the location of nodes in the network. Various network properties, either at the individual-level, sub-group-level, or network-level, describe networks and all are based in some way on network position.[3] *Centrality* is an individual/node-level property of graphs and describes how central each node is in the graph, where central can be defined in various ways. According to Brandes, "it is important to know the relative structural prominence of nodes or links to identify key elements in the network."[5] As described by Freeman, there are three basic measures of centrality: closeness, betweenness, and degree.[6] In order to compare graphs of differ-

ence orders, all three of these measures can be standardized by the maximum theoretical value obtained to create a relative measure. Since centrality is a node-level measure, it is calculated for each node in a graph.

On the other hand, *centralization* considers the effect each node has on the entire graph. Centralization is a global/graph-level property and describes how strongly one particular node influences the rest of the graph. As such, Høivik and Gleditsch stated that "A centralization parameter should ... be a measure of the dispersion in the set of [node] centralities."[7] Similarly, Freeman states that a measure of graph centralization should "index the tendency of a single point to be more central than all other points in the network."[6] He suggests that centralization should use the differences between the most highly centralized node in the graph and the remaining nodes in the graph. The result should be standardized by dividing by the maximum possible centralization value. Because centralization is based on all of the node centralities (which are based on location of a node in the graph), it is directly related to the structure of the graph. The concepts of closeness, betweenness, and degree have been extended to centralization by Freeman.[6] Centralization is a graph-level property and is calculated once for each graph.

It has been shown that the maximum centrality value for closeness, betweenness, and degree is obtained by the center node in a star graph (see Figure 1.1 for a star graph example).[6] It has also been shown that the highest centralization value for closeness, betweenness, and degree are obtained by a star graph.[6, 8]. It should be noted that the definitions of relative centrality and centralization discussed below are designed for static graphs and are not applicable to (non-projected) bipartite graphs or dynamic graphs. However, the theoretical maximum for closeness, betweenness, and degree for each node type in bipartite graphs has been determined.[8, 9] In this way, Freeman's measures can be extended to both node types in bipartite graphs although that is not considered in this research. Freeman thoroughly reviews the historical precedents for these three measures in his 1978 paper so it is not discussed here.[6]

### 1.3.1 Closeness Centrality

One measure of the centrality of a node in a graph is *closeness*. Intuitively, closeness between two nodes is simply a count of how many edges connect one particular node to another node in the graph. In this light, the closest distance between any two nodes in a graph is one, which corresponds to the edge between two adjacent nodes. Large closeness values would be obtained by the two end nodes in a line graph. Since nodes can form edges with more than one other node in a graph, closeness centrality for a particular node is simply the average of all the pair-wise edge counts. Because multiple paths can connect pairs of nodes, the geodesic is chosen for the closeness calculation. By definition, closeness centrality can only be computed for a connected graph (where each node is connected by at least one edge).

Let $d(n_i, n_j)$ be the number of edges in the geodesic between node $n_i$ and node $n_j$. By convention, if node $n_i$ and node $n_j$ are not connected by any edges (meaning they are in different components of the graph), then $d(n_i, n_j) = \infty$. However, a common substitution for $d(n_i, n_j)$ when nodes $n_i$ and $n_j$ are disconnected is $n$.[10, 11] This value is greater than the maximum possible value of $n - 1$, which occurs if the nodes are separated by the maximum possible number of edges. Another solution is to calculate geodesics, and closeness centrality, separately for each component of a disconnected graph. Of course, the number of edges between node $n_i$ and itself is 0, so $d(n_i, n_i) = 0$.

Then the relative closeness centrality (called *closeness centrality* from here forward) of node $n_i$ in an unweighted connected graph is defined as $C'_c(n_i) = \left[\frac{\sum_{j=1}^{n} d(n_i,n_j)}{n-1}\right]^{-1} = \frac{n-1}{\sum_{j=1}^{n} d(n_i,n_j)}$. Note that closeness centrality is the inverse of the average number of edges between node $n_i$ and all of the other nodes in the graph. Freeman shows that the maximum closeness centrality measure is obtained by the central node in a star graph, which is adjacent to every node in the graph.[6] Thus, another interpretation for closeness centrality is the inverse of the ratio by which node $n_i$ is more than one edge away from each of the other nodes.[6]

Closeness centrality is a global measure of node centrality since it is based on geodesics.[12]

Closeness centrality is within the "reachability" class of centrality measures because only nodes in connected graphs can "reach" each other (i.e., are connected).[10] In a communication network, closeness centrality is a measure of information source independence/access or efficiency of information spread. So people with high closeness centrality can avoid the information control of any particular person by simply getting information from another nearby person they communicate with. Also, if viewed in terms of information dissemination, a person with high closeness centrality can quickly spread information through the entire network with the smallest number of people relaying the message.[6, 1] Closeness centrality could also be interpreted as the expected arrival time of information or something that flows through the network. In this way, nodes with high relative closeness receive the information earlier.[13] Borgatti suggests that closeness centrality can be used to model transfer, such as package delivery.[13] Relative closeness centrality was described by Beauchamp in 1965.[14]

A limitation of closeness centrality is that the distribution of values can be limited, especially as graph order increases. The reason is that closeness centrality is based on geodesics and generally, the length of geodesics only increases logarithmically with increasing order of the network. Consequently, the difference between some of the closeness centrality values obtained for nodes in a large graph can be very small.[1] Another limitation is that relative closeness centrality is only defined for connected graphs, although modifications (see the commonly used imputation described above) allow it to be calculated for disconnected graphs as well. Another limitation is that computing all of the geodesics of a network is computationally intense, especially as network order increases.

A version of closeness centrality, *point index centrality*, is described in the 1963 book *Applications of Graph Theory to Group Structure*. In this interpretation, closeness centrality of a node is the sum of all the geodesics of the graph divided by sum of the geodesics for that node. It is not defined for graphs with disconnected nodes.[15]

### 1.3.2 Closeness Centralization

Using Freeman's definition of centralization described above, the equation for relative closeness centralization for graph $g$ (called *closeness centralization* from here forward) is: $C_c(g) = \frac{\sum_{i=1}^{n}[C'_c(n^*)-C'_c(n_i)]}{(n^2-3n+2)/(2n-3)}$, where $C'_c(n^*)$ is the maximum closeness centrality of the observed graph. For closeness centralization, the range of possible values is 0 (a complete graph, a ring graph, or any other graph where all nodes are equivalent) to 1 (a star graph, which has one node that is much more central and all other nodes are all equivalent).

A version of closeness centralization, called *index of centrality* is described by Flament.[15] Index of centrality is computed as the sum of the point index centrality of all of the nodes. It is shown that the minimum index of centrality is obtained by graphs where all nodes are equal. The index of centrality is described as measuring the degree of disparity between the points of a graph.[15] Another version of centralization is average closeness centrality, which is the average of all of the node relative centralities for the graph.[14]

### 1.3.3 Betweenness Centrality

Another measure of centrality of a node is betweenness. Any nodes on the geodesic connecting two nodes are said to be *between* them. Recall that geodesics are not unique and so there may be multiple geodesics that connect a pair of nodes. Betweenness centrality measures how often a node is between other nodes on those geodesics. Betweenness centrality of node $n_k$ is defined as: $C_B(n_k) = \sum_{i}^{n}\sum_{<j}^{n} b_{ij}(n_k)$ where $b_{ij}(n_k) = \frac{g_{ij}(n_k)}{g_{ij}}$ is the number of geodesics connecting $n_i$ and $n_j$ that contain $n_k$ and $g_{ij}$ is the total number of geodesics connecting $n_i$ and $n_j$. Thus, $b_{ij}(n_k)$ is a proportion. If $n_i$ and $n_j$ are in different components, then $g_{ij} = 0$ and $g_{ij}(n_k)$=0, and define $b_{ij}(n_k)$=0. Additionally, if $n_k$ is in a different component than nodes $n_i$ and $n_j$, then $g_{ij}(n_k) = 0$.

It has been shown that the center node in a star graph achieves the maximum possible value of betweenness for a graph with $n$ nodes: $\frac{n^2-3n+2}{2}$.[16] Then, relative betweenness centrality (called from here forward *betweenness centrality*) is defined as: $C'_B(n_k) = \frac{2*C_B(n_k)}{n^2-3n+2}$. Betweenness centrality values range from 0 (disconnected node, peripheral node, or maximally

connected node in a graph) to 1 (center node in star graph). An alternative standardization for betweenness centrality is to divide by $n^2$, the total number of node pairs in the network.[1] Unlike closeness centrality, betweenness centrality can be computed on connected and disconnected networks with no difficulty.

Betweenness centrality of a node has been shown to be equal to the *cutting number* of the node plus the betweenness contribution due to being a bridge.[8] The cutting number of node $n_i$ is the number of unique node pairs for which all paths between the nodes include node $n_i$.[8] A *bridge* is an edge whose removal disconnects the graph.[12] When nodes with high betweenness centrality are removed, the distance between the remaining nodes increases.[17] The betweenness centrality of the center node of a star graph has been shown to decrease by adding edges to the graph.[18] Any node that acts a bridge between two or more groups of nodes will have high betweenness, even if has low degree and it's immediate neighbors have a low degrees.[1] Betweenness centrality can also be extended to directed networks.[1] The bounds of betweenness centrality for a graph are determined by the eigenvalues of the Laplacian matrix of the graph.[19]

Like closeness centrality, betweenness centrality is a global measure since it is based on geodesics.[12] Since it uses geodesics, betweenness centrality is classified as a shortest paths centrality measure.[10] In terms of communication networks, nodes that are between other nodes can distort or control information transfer and so are important for transmission.[6] Betweenness is a measure of the volume of information flowing through a given node from each node to every other node in the network.[13] Removal of nodes with high betweenness from the network can cause a large disruption in transmission of information in the network.[1] When information does not travel along geodesics or the amount of information transmitted between node pairs is unequal, betweenness centrality is not appropriate, although it may provide an approximation of information control.[1] Borgatti suggests that like closeness centrality, betweenness centrality can be used to model transfer, such as package delivery.[13]

Also like closeness centrality, a limitation of betweenness centrality is that it requires computing all of the geodesics of a network. As a result, it is computationally intense, especially

as network order increases.

### 1.3.4 Betweenness Centralization

Relative betweenness centralization of graph $g$ (*betweenness centralization*) is calculated as: $C_B(g) = \frac{\sum_{i=1}^{n} [C'_B(n^*) - C'_B(n_i)]}{n-1}$ where $C'_B(n^*)$ is the maximum betweenness centrality of the observed graph. Like closeness centralization, the range of possible values for betweenness centralization is 0 (a complete graph, a ring graph, or any graph where all nodes are equivalent) to 1 (a star graph, which has only one node that is between all of the others).

### 1.3.5 Degree Centrality

The last measure of centrality described by Freeman is *degree.*[6] The number of edges connected to node $n_i$ is the degree of the node, $k_i$. An equivalent definition of degree of node uses the graph's adjacency matrix instead: $k_i = \sum_{j=1}^{n} A_{ij}$. Equivalently, degree is a count of the number of paths of length one from a given node.[13] Sometimes degree is also called the connectivity of a node. *Degree distribution* is the distribution of values for degree for all nodes in the network.

Degree centrality is defined as: $C_D(n_k) = \sum_{i=1}^{n} a(n_i, n_k)$ and relative degree centrality (*degree centrality*) is defined as: $C'_D(n_k) = \frac{\sum_{i=1}^{n} a(n_i, n_k)}{n-1}$ where

$$a(n_i, n_k) = \begin{cases} 1, & \text{iff edge between } n_i \text{ and } n_k \\ 0, & \text{otherwise} \end{cases}$$

Thus, the numerator of the degree centrality equation is simply the degree count for each node. The maximum possible degree in a graph with $n$ nodes is $n-1$. The range for degree centrality is 0 (disconnected node) to 1 (maximally connected node). Degree centrality is computed the same way for connected and disconnected graphs.

Degree centrality is a local measure since it is based on the immediate contacts of a node.[12] Degree centrality measures how active a node is in the network, meaning how many other

nodes with which it is in direct contact. Since it counts the direct neighbors of a node, degree centrality is also a measure of immediate influence between nodes.[13] Nodes with small degree are more removed from the graph than nodes with larger degree. Like closeness centrality, degree centrality is within the "reachability" class of centrality measures since it directly counts the "reach", or connections, of a node.[10] In terms of communication networks, degree centrality describes the potential communication activity of a node or the number of direct contacts. Borgatti suggests that degree centrality can be used to model money exchange and infections.[13] Degree centrality was first described by Nieminen in 1974.[20]

Degree centrality is a count of all edges connected to a node, thereby assuming an undirected network. In the case of a directed network, *in-degree* and *out-degree* are used instead. In-degree is a count of the number of edges leading to a node while out-degree is the number of edges leaving a node.

### 1.3.6 Degree centralization

Relative degree centralization of graph $g$ (*degree centralization*) is calculated as: $C_D(g) = \frac{\sum_{i=1}^{n} [C_D(n^*) - C_D(n_i)]}{n^2 - 3n + 2}$ where $C_D(n^*)$ is the maximum degree centrality of the observed graph. Again, the range of possible values is 0 to 1. The exact bounds of degree centralization by graph density and graph order has been examined.[21] Degree centralization is also called connectivity centralization.

## 1.4 Conclusion

Network science provides a new framework to examine important topics in public health. However, in order to make use of network science concepts and applications, network terminology and measures must be defined. The goal of this chapter has been to describe the required network vocabulary along with the centrality and centralization measures that will be used throughout this dissertation. Although there are limitations to the Freeman

measures of closeness, betweenness, and degree centrality, they are still widely used in the literature and are often the gold standard to which new centrality measures are compared.

# Chapter 2

# Literature Review

## 2.1 Introduction

Much research has been conducted on the idea of centrality in networks with topics ranging from the development of new measures, computational algorithms to expeditiously compute geodesics, and problems in the application of centrality to sampled networks. Additional research has produced network generating methods designed to generate graphs with particular properties. Unfortunately, very little literature has been published on centralization. The following is a very brief literature review around the ideas of centrality and centralization, with a special focus on the Freeman version of these measures. Refer to Chapter 1 for definitions of graph terms and an explanation of the Freeman centrality and centralization measures.

## 2.2 Relative Centrality/Centralization Properties

According to Dwyer, "centrality analysis determines the importance of vertices in a network based on their connectivity within the network structure."[22] As described in the previous chapter, closeness, betweenness, and degree centrality, although all node-level measures, describe different attributes of the same network. The relationship between these measures

has been investigated along with how they are influenced by sampling error.

**Relative Centrality Properties** Valente et al. examine the correlation between nine centrality measures (symmetrized degree, in-degree, out-degree, symmetrized betweenness, betweenness, symmetrized closeness, closeness-in, closeness-out, and eigenvector centrality) collected from seven studies with 58 networks. They calculated the average correlation, standard deviation, and range across centrality measures. The study found for the undirected measures a very strong correlation between degree centrality and eigenvector centrality ($r = 0.92$), a moderately strong correlation between degree centrality and betweenness centrality ($r = 0.71$), between degree centrality and closeness centrality ($r = 0.66$), between betweenness centrality and eigenvector centrality ($r = 0.64$), and between closeness centrality and eigenvector centrality ($r = 0.64$). They also found a weak correlation between closeness centrality and betweenness centrality ($r = 0.37$).[23]

A study of the correlation between degree and betweenness centrality found a strong positive correlation when the network had a scale-free degree distribution (a small number of nodes with very large degree and many nodes with small degree) and was much weaker in other types of networks.[24] In fact, it has been shown that there is a positive correlation between degree centrality and betweenness centrality on real social networks.[25] The betweenness centrality of a node is related to its degree and the exponent of the power law degree distribution in Barabási-Albert Preferential Attachment graphs.[26] Generally, for scale-free graphs the betweenness centrality distribution follows a power law.[26]

**Relative Centralization Properties** Nakao describes the distribution of relative centralization values in small, connected networks (5-8 nodes).[27] Also described in the paper is the maximum centralization value obtained when the number of edges is set and the correlation between centralization pairs. Another paper uses simulation to produce a null distribution for betweenness and degree centralization conditional on graph density.[28]

**Sampling Networks** When trying to capture the structure of a real network, sampling becomes an important issue. Costenbader and Valente use network data collected from eight different studies (creating 59 networks) to determine the correlation between 11 different centrality measures (in-degree, out-degree, degree symmetrized, betweenness directed, betweenness symmetrized, closeness directed, closeness symmetrized, first eigenvector/simple eigenvector, eigenvector centrality, radiality, and integration) obtained from the full network and a sampled network using sampling percentages of 10% to 80% by 10% increments. Results show that correlation decreases as the sampling proportion decreases. In-degree retained the greatest correlation while the remaining measures declined fairly steeply and eigenvector centrality was wave-like.[29]

An examination of the robustness of closeness, betweenness, degree, and eigenvector centrality to different types of sampling error (such as node removal, node addition, edge addition, or edge deletion) using random graphs of different orders and density found that the measures behave similarly in accuracy (as measured by proportion of nodes ranked similarly between original network and the sampled networks). The authors also found that node errors reduced accuracy less than edge errors and that for all measures the percentage of errors increased as the accuracy decreased.[30] A related study extended the investigation of the robustness of those centrality measures to clique, core/periphery, and preferential attachment networks and found the results varied by network type.[31] Core/periphery graphs were very sensitive to edge removal and node addition, clique graphs were sensitive to node and edge addition, preferential attachment graphs were the least sensitive to all types of error. The reliability of directed versions of closeness, betweenness, degree, and flow betweenness (a measure of the maximum possible flow over all paths between a node pair- see Section 2.3 Other Centrality Measures) was evaluating using a study of social support of high school students. The students were surveyed three times in short succession about their relationships and the correlation between responses for each measure calculated. In-measures had a higher correlation than out-measures with degree and closeness performing very well and betweenness and flow betweenness performing more poorly.[32]

## 2.3 Other Centrality Measures

This research focuses on Freeman's relative centrality measures but many other centrality measures have been described in the literature. In fact, some of the other centrality measures are even based on the Freeman measures of centrality. For example, relative closeness centrality has been used as basis for other measures, such as bridging.[12] Interestingly, Valente and Fujimoto also use a star graph to standardize their measure of bridging.[12] A modified degree centrality, called local centrality, considers the nearest and next-nearest neighbors of a given node.[33] In order to find the nodes that are the best performers overall, all-around nodes have been defined as nodes that have high centrality for betweenness, degree, and k-core.[25]

Using the process model of social influence, Friedkin developed three new centrality measures: total effects centrality, immediate effects centrality, and meditative effects centrality.[34] Combine Centrality Actor Ranking (CCR) is the sum of the (non-relative) degree, betweenness, and closeness centralities for a particular node. This was used in identifying leaders in a terrorist network.[35] Gil Schmidt power centrality index is a weighted sum of the $k$-th neighbors of a node.[36] The Gil Schmidt power centrality index was developed to find important political actors in Mexico.[36] Second order centrality is based on the standard deviation of sum of the lengths of time until a specified node is encountered on a random walk through the network.[37] Second order centrality can be calculated distributively, thus saving computing time for large networks. Subgraph centrality measures node participation in subgraphs of the main graph by weighting the number of closed walks starting and ending at the same node by the length of the walk.[38] Control centrality in directed weighted networks describes the control of each node in the network based on the generic rank of a controllability matrix.[39] Eccentricity is the longest geodesic between node $n_i$ and any of the other nodes in the network. In other words, the eccentricity of $n_i$ is $\max d(n_i, n_j), \forall j \in V(g)$.[10]

Measures of centrality that use current flow, where information travels through all paths in the network instead of just the shortests paths, include flow betweenness and flow close-

ness.[5] Flow betweenness calculates the maximum possible flow, overall all possible path choices, between nodes.[1] Unlike flow betweenness, random walk betweenness incorporates absorbing random walks between node pairs so it includes all possible paths, not just the paths that produce maximum flow of information.[40, 1] This modification represents no prior knowledge of the shortest paths to send information from one node to another. A slight modification of random walk betweenness, allows nodes to be repeated on the walk.[41] Information centrality calculates all possible paths between node pairs (not just the geodesics).[42] It has been shown that flow closeness is the same as information centrality.[5]

The efficiency of information flow between nodes $n_i$ and $n_j$ is $\epsilon_{ij} = \frac{1}{d(ij)}$. Then, the local efficiency of node $n_i$, also called point closeness, is $E(n_i) = \frac{1}{n}\sum_{j \neq i} \frac{1}{d(ij)}$.[43, 44] Local efficiency has been used to describe fault tolerance of a network, or how robust the network is in terms of information flow when node $n_i$ is removed. It is also a relative measure so can be compared between graphs of different orders. Efficiency is related to closeness centrality by inverting the geodesic between all node pairs. Another alternative definition of closeness of node $n_i$ is defined as $C(n_i) = \sum_{j \neq i} \frac{1}{2^{d(ij)}}$.[44] Let $d_k(ij)$ be the geodesic between nodes $n_i$ and $n_j$ when node $n_k$ and all of its links are deleted. Then residual closeness of node $n_k$ is defined as $C(n_k) = \sum_i \sum_{j \neq i} \frac{1}{2^{d_k(ij)}}$.[44] Residual closeness can be interpreted as a measure of robustness of a network. Formulas for residual closeness values for star, wheel, and complete graphs have been determined.[45]

The centrality measures previously discussed are node-level measures. Edge-level centralities can also be calculated. An example of an edge centrality is amongness centrality. It is calculated in a weighted, directed graph and is based on a function of the edge weights.[46] Edge betweenness has also been defined.[19]

### 2.3.1 Eigenvector Centrality

In addition to closeness, betweenness, and degree, eigenvector is another commonly used centrality measure. Eigenvector centrality is the principal eigenvector of a undirected con-

nected graph.[47] Let $A$ be the adjacency matrix of graph $g$ and $\lambda$ be the first eigenvalue. Then $\lambda x = Ax$. So, $x_i = \lambda^{-1} \sum_j A_{ij} x_j$. This means that the eigenvector centrality of a node is proportional to the sum of the centralities of the nodes that it is connected to. As such it can be considered as a popularity measure describing how connected a node is to other well-connected nodes. Thus, it is a measure of node influence because it incorporates not only the direct contacts of a node but also the contacts of those contacts and the contacts of those contacts and so on. Another interpretation of eigenvector centrality is a summary of the number of walks of any length, weighted inversely by walk length, from a given node.[13] A node may have high eigenvector centrality if it is has a high degree or is connected to neighbors with a high degree or both.[1] If a graph has several components, eigenvector centrality must be calculated for each component separately.[47, 48] Eigenvector centrality is always non-negative.[1] Ruhnau proved when normalizing the eigenvector with its Euclidean norm, the maximum value is obtained only by the center node in a star graph.[49] Bonacich extended eigenvector centrality to bipartite graphs and a later correction adjusts this score by the group size.[50, 51] Bonacich also created a similar measure called power or beta centrality that can be used for both power and bargaining relationships.[52] Two separate measures developed by Bonacich and Lloyd provide eigenvector centrality for directed networks.[53]

## 2.4 Other Centralization Measures

Again, this research focuses on Freeman's centralization measures but several others have been described and examined in the literature. For example, Tallberg looked at centralization measures such as maximum centrality minus average centrality, variance of centrality measures, and maximum centrality. These measures were not standardized by maximum theoretical value.[54] Others have considered average betweenness and maximum betweenness for betweenness centralization.[19, 55] Centralization has also been considered for the Gil Schmidt power centrality index and is defined similarly to the Freeman centralization measures.[56] Interestingly, like the Freeman relative centralization measures, the maximum

Gil Schmidt power index centralization is proved to be obtained by a star graph.

In-degree and out-degree variance were used as centralization measures in an analysis of a version of a trust game with a variable incentive for sellers to abuse a buyer's trust.[57] A measure utilizing the variance in degree centrality was also developed as an "index of heterogeneity" for graphs.[58] A global measure of the efficiency of information flow in a graph is defined as $E(g) = \frac{1}{n(n-1)} \sum_{i \neq j \in g} \frac{1}{d(ij)}$[43] Global efficiency is normalized to produce a value between 0 and 1. Based on the alternative definition of closeness of node $n_i$, where closeness is defined as $C(n_i) = \sum_{j \neq i} \frac{1}{2^{d(ij)}}$, a global measure of closeness for graph $g$ can be calculated: $C = \sum_i C(n_i)$.[44] Using residual closeness, vertex residual closeness is $R = min_k C_k$. Vertex residual closeness can be normalized by dividing by the original graph closeness $C$.

## 2.5 Networks in Public Health

Network analysis in public health is becoming more common. It is not surprising given how useful network science can be in the investigation of the relationships between individuals, an important factor in many public health issues. A recent review article suggests that researchers interested in three categories of public health networks use network analysis: transmission networks (either information or disease), social networks, and organizational networks.[3] For example, smoking behavior in a large network was examined by Christakis and Fowler.[59] Network analysis was used to examine an outbreak of gonorrhea in Canada and to describe the structure of an adolescent sexual network.[60, 61] Another study examined the spread of hospital-associated infections to patients by healthcare workers.[62] At an organizational level, the structure of state tobacco control networks was investigated.[63]

Network analysis methods have been applied to further examine real-life networks on infectious diseases. A study investigating the sexual network of adolescents found the structure of the network important in how disease was transmitted.[61] For example, network analysis of a gonorrhea outbreak in Canada suggested that important individuals in propagating

the disease could be identified by information centrality.[60] Network analysis has been used to evaluate the Severe Acute Respiratory Syndrome (SARS) outbreak and possible public health interventions.[64, 65]

Within the studies that use network methods, a wide variety of network measures were reported. In the Canadian gonorrhea outbreak, network density, degree centrality (not relative degree centrality), betweenness centrality (not relative betweenness centrality), information centrality, and 2-core membership were reported.[60] The study of the sexual network of adolescents utilized Bonacich power centrality.[61]

**Relative Centrality in Real Networks** Relative betweenness centrality was used as a measure of contact frequency in an analysis of state tobacco control networks.[63] Degree, closeness, eccentricity, eigenvector, betweenness, random walk betweenness, and second order centrality were compared in the largest connected component of the jazz players collaboration network. The authors found that the top ten nodes for all measures excepting eccentricity generally agreed. They also stated that closeness and degree centrality produce a stronger correlation and eccentricity results are dissimilar to second order centrality results.[66]

**Relative Centralization in Real Networks** Centralization is reported for the Padgett's Florentine Families network of marital relations.[56] Freeman's degree and betweenness centralization measures were used to help identify linking-pin organizations among a group of organizations involved in providing assistance during natural disasters.[67] Betweenness centralization was used to describe the network of organizations that responded to the 1989 Exxon Valdez oil spill.[68] Betweenness centralization was used to compare state tobacco control programs.[63]

## 2.6 Network Programs

**Programs that Produce Relative Centrality/Centralization** Pajek computes closeness, betweenness, and degree centralization as well as the centrality measures for closeness, betweenness, and degree.[69] Freeman's measures are used to compute closeness, betweenness, and degree centralization. Note that closeness centrality in Pajek is only computed for that fraction of nodes that are connected in network. In this case, a modified closeness centrality for connected nodes is reported: weight the sum of the geodesics for each node by the percentage of nodes that are connected. Disconnected nodes are defined to have a closeness centrality of 0. Closeness centralization is not calculated in the case of disconnected nodes.[70]

UCINET computes Freeman's relative closeness, betweenness, and degree centrality and centralization measures along with some variations of them.[71]. Note that UCINET only calculates closeness for connected graphs. The *igraph* package in R computes relative closeness, betweenness, and degree centrality and centralization for all graphs (connected or otherwise), although there are some errors in the calculations.[72, 73] Note that *igraph* uses a simple imputation of $\frac{1}{n}$ for the average path length of a disconnected node.[10, 11] This value is then used in the closeness calculations for disconnected graphs. Gephi, a free network analysis and graphing program, computes relative closeness and betweenness centrality but not relative degree centrality or any relative centralization measures.[74] NodeXL, another popular free network analysis program, also does not compute any relative centralization measures.[75]

**Algorithms for Centrality** In order to calculate closeness centrality and betweenness centrality, the geodesic between all pairs of node in a graph must be calculated (the all-pairs shortest-paths, or APSP, problem). This is a computationally intensive problem, especially for large networks. The original algorithm for calculating all the geodesics of a network of any density requires $O(n^3)$ computation time.[76, 33] When the network is sparse, an alternative algorithm only requires $O(n^2 log n + nm)$ computation time.[77, 78, 33] Lastly,

Brandes created an even faster algorithm for betweenness centrality that is implemented in the *igraph* package.[79] For sparse networks, it requires $O(nm)$ computation time.[33, 79] In all of these calculations, $n$ is the number of nodes and $m$ is the number of edges.

There are other algorithms that are also optimized for sparse networks.[80] Kanchi and Vineyard describe a distributed algorithm to solve the APSP problem by partitioning the network.[81] Yet another study developed a method to compute centrality using Graphical Processing Units instead of the traditional Central Processing Unit (CPU) method for faster computation.[82] Brandes also developed a method to estimate closeness and betweenness for very large networks based on randomly sampling nodes in the graph.[83]

In a completely different approach to the APSP problem, modifications to the centrality measures themselves have been considered. Many different changes to the centrality algorithms, each with different attributes and failings, have been proposed as solutions and work continues to be done in this area. For example, there is an algorithm that ranks the top closeness centrality values for large networks that is faster than programs that actually obtain the values.[84] Another method approximates closeness centrality for very large networks that is also faster than programs that actually obtain the values.[85]

**Visualizing Networks** GEOMI (Geometry for Maximum Insight) program provides a way to visualize all of the centrality values for a network up to 100 nodes, although the software seems to have been replaced by a visualization program designed for gene expression data.[22] Additionally, all of the commonly used network programs described above (Pajek, UCINET, *igraph*, etc.) provide simple visualization tools.[69, 75, 72, 74, 71] Gephi is a particularly nice graph visualization tool.[74]

## 2.7 Generating Networks

Many different algorithms have been developed to generate networks with specific structures. Popular models include small world models, which generate networks with short path lengths and high clustering and Barabàsi-Albert preferential attachment models which

generate networks with a power-law degree distribution. Many of the network software programs discussed earlier implement versions of these algorithms.

**Erdös-Rènyi Random Graphs** Erdös-Rènyi methods to produce random graphs are the classic way to create graphs that do not emphasize any particular structure.[86] In the $Gnm$ version of the Erdös-Rènyi random graph, a specified number of edges are randomly added to a network of a particular order. More specifically, $m$ edges are drawn uniformly randomly from the set of all possible edges and added to an empty graph.

**Barabási-Albert Preferential Attachment** The Barabási-Albert (BA) preferential attachment model is a common method of generating networks with a power law degree distribution.[87, 88] Preferential attachment models attempt to reproduce the "rich get richer" phenomenon observed in many systems. Examples of networks with a power law degree distribution are collaborations between movie actors, the world wide web, and scientific citations.[87, 88] Graphs with a power law degree distribution have a small number of nodes with very large degree and many nodes with small degree and so are also called scale free. In the BA model, new nodes are added to a graph and the links from the new nodes preferentially bond to the established nodes with the most connections. Modifications to the BA preferential attachment model include allowing new links to form between any nodes (new or established) in the network.[89]

**Small World Graphs** Small world graphs are designed so that the clustering coefficient (equal to the observed number of triangles over number of triples in the graph) is high while the average path length is simultaneously low. In general, the mean geodesic of a small world graph increases as the log of the graph's order increases.[1] In this method, a circular lattice graph where $k$ neighbors are linked by edges is systematically rewired with probability $p$ of an edge forming between two nodes.[90]

**Two Node Type Preferential Attachment** Heterogeneity in preference for forming edges between nodes of different types is used to understand the importance of node com-

munities in a larger network and is the basis for homophily, or the idea that similar nodes will link to each other.[91] In this method, nodes are randomly divided into a set number of types that link to each other with different probabilities. Most versions of node type preferential attachment add nodes and edges over time.[92, 93, 91]

**Other Methods** SpecNet is a network generating program that uses a spatial network algorithm to create networks that exhibit certain properties. Based on network order (number of nodes) and mean degree, SpecNet can generate networks with specific assortativity, clustering, and/or fragmentation. Importantly, this program does allow for disconnected nodes.[94]

## 2.8 Conclusion

The network science published literature is too vast to consider without selecting a subset of interest. This has been a very brief review of the network literature focusing on the idea of centrality and centralization. One rapidly expanding area of research not previously discussed is dynamic networks. Common methods of handling dynamic graphs involve condensing the information into one static graph and applying traditional network methods. More recently, centrality has been re-defined for dynamic graphs.[95]

# Part II

# Topic 1

# Chapter 3

# Centralization in Various Graph Generating Methods

## 3.1 Introduction

Centralization is a measure of how important one node is compared to all of the other nodes in a network. Unlike centrality, which is a node-level measure, centralization is a graph-level measure and so is measured once for each network. According to Freeman, centralization should "index the tendency of a single point to be more central than all other points in the network."[6] This measure could be useful for public health applications because it might indicate how susceptible the network is to interventions that are implemented at the most important node. Indeed, previous research using a network application of the Diffusion of Innovations theory has shown that the centrality of adopters influences how information moves through the network.[3] In communities, "opinion leaders are people who influence the opinions, attitudes, beliefs, motivations, and behaviors of others."[96] The most central node in a very centralized network could be an opinion leader whose recommendations are very influential to the followers in the network. However, very little work has been done investigating the basic properties of centralization so these areas must first be explored before applied research can be conducted.

Currently, no research describes the centralization values produced by common network generating methods. As a result it is unclear which graph generating method should be employed in order to investigate the properties of centralization. Therefore, the primary goal of this paper is to describe the full range of centralization values produced by the common network generating methods in Erdös-Rènyi random graphs, Barabási-Albert preferential attachment graphs, small world graphs, and two node type preferential attachment graphs as well as a new method call Star Start. Due to computational limitations, small networks of 5-20 nodes will be investigated. Additional goals of this research are to determine the frequency with which centralization values are obtained and the range of graph structures that produce the centralization values for each method. Note that this study does not attempt to characterize the distribution of centralization values for each graph generating method as the distributions are likely different. Instead, this study aims to describe the coverage of centralization values by each method.

### 3.1.1 Measures

For this paper we will call a particular network a *graph*, *nodes* are the points in the graph and *edges* connect the nodes. The number of edges in a graph is called the graph's *size* while the number of nodes in a graph is called the graph's *order*. A *path* is the sequence of edges between two nodes. The *geodesic* is the shortest path between two nodes, where distance is defined as the count of number of edges.

A graph is *connected* if there is a path between each node in the graph and *disconnected* otherwise. Graph *isomorphisms* refer to graphs that have the same structure although the edges creating the graph are different. A graph is *simple* if there is only one edge connecting nodes (multiple or weighted edges are not allowed). A *digraph* is a directed graph, where the edges are directed from one node to another. Otherwise, edges are *undirected*. *Self-loops* are edges that start and end with the same node.

This paper uses the relative centrality and centralization measures for closeness, betweenness, and degree originally described by Freeman to facilitate comparisons between graphs of

different orders.[6] Relative measures are obtained by taking the observed value for a particular graph and dividing by the theoretical maximum value for that measure. Using relative measures, the range of possible values for both centrality and centralization fall between zero and one. For closeness, betweenness, and degree centrality, the theoretical maximum has been proven to be achieved by the center node of a star graph.[6] For closeness, betweenness, and degree centralization the theoretical maximum has been proven to be achieved by the star graph.[6, 8] A brief description of each of these measures is shown below. These measures as defined below apply to simple, undirected, and unweighted graphs.

**Closeness** Intuitively, closeness between two nodes is simply a count of how many edges connect one particular node to another node in the graph. In this light, the closest distance between any two nodes in a graph is one, which corresponds to the edge between two adjacent nodes. Let $d(n_i, n_j)$ be the number of edges in the geodesic between node $n_i$ and node $n_j$. By convention, if node $n_i$ and node $n_j$ are not connected by any edges, then $d(n_i, n_j) = \infty$. Of course, the number of edges between node $n_i$ and itself is 0, so $d(n_i, n_i) = 0$. Then the relative closeness centrality (called *closeness centrality* from here forward) of node $n_i$ (in a connected graph) is defined as $C'_c(n_i) = \left[\frac{\sum_{j=1}^{n} d(n_i,n_j)}{n-1}\right]^{-1} = \frac{n-1}{\sum_{j=1}^{n} d(n_i,n_j)}$. By definition, closeness centrality can only be computed for a connected graph (where each node is connected by at least one edge). However, *igraph* package in R substitutes $n$ for $\infty$ in the case of disconnected nodes so all nodes have closeness centrality values.[10, 11] Consequently, for this paper closeness centrality is computed for all graphs, connected or disconnected.

Relative closeness centralization (called *closeness centralization* from here forward) is: $C_c = \frac{\sum_{i=1}^{n}[C'_c(n^*)-C'_c(n_i)]}{(n^2-3n+2)/(2n-3)}$, where $C'_c(n^*)$ is the maximum closeness centrality of the observed graph. Given this formula, it is clear that a centralization of 0 is obtained by a graph where all nodes are equal (ex: ring, empty, or complete graph).

**Betweenness** Any nodes on the geodesic connecting two nodes are said to be between them. Betweenness centrality measures how often a node is between other nodes. $C_B(n_k) =$

$\sum_{i}^{n}\sum_{<j}^{n} b_{ij}(n_k)$ where $b_{ij}(n_k) = \frac{g_{ij}(n_k)}{g_{ij}}$ is the number of geodesics connecting $n_i$ and $n_j$ that contain $n_k$ and $g_{ij}$ is the total number of geodesics connecting $n_i$ and $n_j$. Since the center point in a star graph obtains the maximum value, relative betweenness centrality (*betweenness centrality*) is defined as: $C'_B(n_k) = \frac{2*C_B(n_k)}{n^2-3n+2}$.

Relative betweenness centralization (*betweenness centralization*) is calculated as: $C_B = \frac{\sum_{i=1}^{n}[C'_B(n^*)-C'_B(n_i)]}{n-1}$ where $C'_B(n^*)$ is the maximum betweenness centrality of the observed graph.

**Degree** The number of edges connected to node $n_i$ is the degree, $k_i$, of the node. Degree centrality is defined as: $C_D(n_k) = \sum_{i=1}^{n} a(n_i, n_k)$ and relative degree centrality (*degree centrality*) is defined as: $C'_D(n_k) = \frac{\sum_{i=1}^{n} a(n_i,n_k)}{n-1}$ where

$$a(n_i, n_k) = \begin{cases} 1, & \text{iff edge between } n_i \text{ and } n_k \\ 0, & \text{otherwise} \end{cases}$$

Relative degree centralization (*degree centralization*) is calculated as:
$C_D = \frac{\sum_{i=1}^{n}[C_D(n^*)-C_D(n_i)]}{n^2-3n+2}$ where $C_D(n^*)$ is the maximum degree centrality of the observed graph.

## 3.2 Methods

### 3.2.1 Erdös-Rènyi Random Graphs

Erdös-Rènyi methods to produce random graphs are the classic way to create graphs that do not emphasize any particular structure.[86] In the *Gnm* version of the Erdös-Rènyi random graph, a specified number of edges are randomly added to a network of a particular order. More specifically, $m$ edges are drawn uniformly randomly from the set of all possible edges and added to an empty graph. Thus, the Erdös-Rènyi *Gnm* method produces the distribution of all possible graphs with $m$ edges selected at random. However, the goal

of this research is to determine the range of centralization values possible for a particular network generating method, so centralization was measured as edges were incrementally added. The number of edges added to the network ran from zero (creating an empty graph) to $\frac{n(n-1)}{2}$ (creating a complete graph), incrementing by one edge. For each graph order, 500 repetitions were made for each number of edges that were randomly added and all three centrality and centralization measures calculated for each graph produced. Note that $Gnm$ graphs are not required to be connected.

### 3.2.2 Barabási-Albert Preferential Attachment

The Barabási-Albert preferential attachment model is a common method of generating networks with a power law degree distribution.[87, 88] Examples of networks with a power law degree distribution are collaborations between movie actors, the world wide web, and scientific citations.[87, 88] Graphs produced by this method have a small number of nodes with very large degree and many nodes with small degree and so are also called scale free.

In the version investigated here, the graph starts with one node and then at each time step one node is added and one edge is added to the graph. To form the new edge, the new node chooses an old node (already in the graph) randomly based on its degree and forms one edge with it. By design, the probability that established node $n_i$ is chosen is proportional to its degree plus a constant: $\mathrm{P}(n_i \text{ chosen}) = k_i^{*\alpha} + a$, where $k_i^*$ is the number of connections $n_i$ has received (e.g., $k-1$), $\alpha$ is the power of preferential attachment, and $a$ is the attractiveness of nodes with no edges. The process of adding one node and one edge is repeated until the order of the graph reaches the specified number of nodes. As a result, Barabási-Albert preferential attachment networks are dynamically generated.

The powers of preferential attachment investigated included 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100. The constant added to the degree of a node followed the same range and every combination of constant and power was tested. Measurements were calculated for each graph and 500 graphs were

produced for each combination of constant and power and all three measures of centrality and centralization calculated. Note that this method produces only connected graphs since each new node is immediately linked to an old node.

### 3.2.3 Small World Graphs

Small world graphs are designed so that the clustering coefficient (equal to the observed number of triangles over number of triples in the graph, where a triangle is 3 nodes connected to each other to form a ring while a triple is simply three nodes that form a component together) is high while the average path length is simultaneously low. In general, the mean geodesic of a small world graph increases as the log of the graph's order increases.[1] In this method, a circular lattice graph where $k$ neighbors are linked by edges is systematically rewired with probability $p$ of an edge forming between two nodes.[90] The value of $p$ ranged from 0 to 1 in increments of 0.01 and the size of the neighborhood ($k$) varied from 0 to $n$, the entire network. For each combination of values, 500 graphs were calculated and the three measures of centrality and centralization calculated. Note that small world graphs are not required to be connected.

### 3.2.4 Two Node Type Preferential Attachment

Heterogeneity in preference for forming edges between nodes of different types is used to understand the importance of node communities in a larger network and is the basis for homophily, or the idea that similar nodes will link to each other.[91] In this method, nodes are randomly divided into a set number of types that link to each other with different probabilities. This paper investigated an equal probability of being either of two node types. The probability of a link forming between two nodes of the same type was $p_1$ and the probability of two nodes of different types forming a link was $p_2$. Both $p_1$ and $p_2$ ranged from 0 to 1 in increments of 0.01 and all combinations of $p_1$ and $p_2$ were tried. For each combination of values, 500 graphs were produced and the three measures of centrality and centralization calculated. Note that two node type preferential attachment graphs are not

required to be connected. Unlike the static version used here, most versions of node type preferential attachment add nodes and edges over time.[92, 93, 91]

### 3.2.5 Star Start

This method starts with a star graph, $g$, with $n$ nodes and then edges were randomly added or deleted from the graph with equal probability until $g$ was complete (i.e., contained the maximum possible number of edges: $|E(g)| = \frac{n(n-1)}{2}$) or empty (contained no edges linking any nodes: $|E(g)| = 0$). Thus, in order to complete an iteration, either a minimum of $n-1$ edges must be removed from the star graph to create an empty graph or a minimum of $(n-1)(n-2)/2$ edges must be added to the star graph to create a complete graph. All three measures of centrality and centralization were taken for each update of $g$. This process was repeated 500 times (iterations), each time starting with a star graph with $n$ vertices and continuing until the graph was complete or empty.

Note that for all of the methods, graphs were undirected and multiple edges and self-loops were not allowed. Graph order ranged from 5 to 20 nodes in increments of one node. All graphs were generated in R version 2.15.1 with the *igraph* package version 0.6-2.[73, 72]

## 3.3 Comparisons

In order to compare the different graph generating methods, centralization values were binned into 101 bins from 0 to 1 in increments of 0.01. Bar charts plotted the presence or absence of values in a particular centralization bin for each method for closeness, betweenness, and degree. Since each of the graph generating methods produced a variable number of graphs with a particular centralization value and primary concern rests with finding what centralization values are commonly produced by these programs rather than the exact number of graphs, ranks were used to compare the relative frequency of graphs in each centralization bin for each method. The exact numbers of graphs that fall into a

particular centralization bin can be modified by finding the right combination of parameters and so the frequency distribution is not used. Centralization bins were ranked by number of graphs such that larger numbers of graphs have larger ranks and then the ranks were scaled to between 0 and 1 to create a relative rank. Relative rank was used so that the ranks could be plotted on the same scale for all graph orders. Since graphs with different structures may have the same centralization value, scatter plots of maximum centrality and centralization were plotted for each method as a means of illustrating the range of graphs that fall into a particular centralization bin. In these plots, opacity of the points reflects the number of maximum centrality values that fell into a particular centralization bin with darker colors meaning larger numbers of graphs. Comparisons were performed for each graph order between 5 and 20 nodes and results summarized by graph order. All computations were performed in R version 2.15.1 with the *igraph* package version 0.6-2.[73, 72]

## 3.4 Results

### 3.4.1 Closeness

**Centralization Values Obtained** Movie 3.1 describes the range of centralization values produced by each of these methods for each graph order for closeness centralization. As shown in the movie, Erdös-Rènyi random graphs (ER) and small world graphs (SW) fail to produce graphs with high levels of closeness centralization except at very small graph orders (five or six nodes). However, over the range of graph orders examined, these graph methods do produce all centralization values between 0 and 0.6 (moderate centralization). On the other hand, Barabási-Albert (BA) preferential attachment graphs do not produce low centralization graphs but do produce graphs with the full range of centralization values from about 0.2 to 0.6. Additionally, the BA method does produce graphs with moderate to high centralization although the range is limited. Lastly, two node type preferential attachment (Pref) and Star Start (Star) produce graphs that cover most of the full theoretical range, with only a few small gaps at higher centralization levels as the graph order increases. Star Start produces a few more unique highly centralized values than the two node type

preferential attachment model. A comparison of just these two methods is shown in the Appendix as Movie 3.10.

**Rank of Centralization Values** Movie 3.2 describes the rank of centralization values produced by each of these methods for each graph order for closeness centralization. As shown in the movie, as graph order increases, a bell-shaped pattern emerges regarding frequency of values obtained for closeness. Most of the methods demonstrate a peak at very low centralization levels (around 0.2). Barabási-Albert preferential attachment models have a slightly higher peak centralization level (around at 0.4). Barabási-Albert preferential attachment models also have high ranks for centralization values between 0.8 and 1, indicating that many of the graphs generated fall into those bins. Star Start has moderate ranks for centralization values between 0.8 and 1 and the other methods all have very low ranks.

**Range of Graph Structures** Movie 3.3 further describes the distribution of closeness centralization values by plotting them against the maximum closeness centrality value for each method. As shown in the movie, two node type preferential attachment and Star Start have the greatest range of maximum centrality values for a particular centralization bin while the range for Barabási-Albert preferential attachment models is very restricted. However, as graph order increases there are larger gaps in the maximum centrality values at very high centralization levels, especially for Star Start. At low centralization levels, Erdös-Rènyi random graphs and small world graphs also have a broad range of maximum centrality values which continues to moderate centralization levels for Erdös-Rènyi random graphs.

### 3.4.2 Betweenness

**Centralization Values Obtained** Movie 3.4 describes the range of centralization values produced by each of the graph generating methods examined for each graph order for betweenness centralization. Erdös-Rènyi random graphs and small world graphs produce all

centralization values between 0 and 0.6 (moderate centralization). Again, Barabási-Albert preferential attachment graphs do not produce low centralization graphs but do produce graphs with the full range of centralization values from about 0.2 to 0.6. Unlike for closeness, the Barabási-Albert preferential attachment method does produce graphs with moderate to high centralization with increasing coverage as graph order increases. Again, two node type preferential attachment and Star Start produce graphs that cover most of the full theoretical range, with Star Start producing slightly more unique highly centralized values. A comparison of just these two methods is shown in the Appendix as Movie 3.11.

**Rank of Centralization Values** Movie 3.5 describes the rank of centralization values produced by each of these methods for each graph order for betweenness centralization. As shown in the movie, as graph order increases, a completely different pattern emerges regarding frequency of values obtained for betweenness. Most of the methods demonstrate a rank peak at 0 (completely decentralized) and rank decreases linearly as centralization increases. Barabási-Albert preferential attachment models are a notable exception, still following a bell-shaped curve that then curves upward to have high ranks for centralization values between 0.8 and 1. Again, Star Start has moderate ranks for centralization values between 0.8 and 1 and the other methods all have very low ranks.

**Range of Graph Structures** Movie 3.6 further describes the distribution of betweenness centralization values by plotting them against the maximum betweenness centrality value for each method. As shown in the movie, most of the methods display a similar range of maximum centrality values for the centralization bins that are obtained. Star Start and Barabási-Albert preferential attachment have the greatest range of maximum centrality values for very high centralization levels even as graph order increases. On the other hand, the maximum centrality range for two node type preferential attachment becomes very restricted as graph order increases. Small world graphs have a slightly wider range of values than Erdös-Rènyi random graphs at low to moderate centralization levels.

### 3.4.3 Degree

**Centralization Values Obtained** Movie 3.7 describes the range of centralization values produced by each of the graph generating methods examined for each graph order for degree centralization. As for closeness Erdös-Rènyi random graphs produce all centralization values at the smallest graph order (five nodes) and this range decreases as the graph order increases ending with a range between 0 and 0.6 (moderate centralization). Small world graphs produce graphs with centralization values between 0 and 0.6 but the range of unique values is very sparse. Barabási-Albert preferential attachment graphs do not produce very low centralization graphs but do produce highly centralized graphs. Similar to the results for closeness and betweenness, two node type preferential attachment and Star Start produce graphs that cover most of the full theoretical range, with Star Start producing slightly more unique highly centralized values. A comparison of just these two methods is shown in the Appendix as Movie 3.12.

**Rank of Centralization Values** Movie 3.8 describes the rank of centralization values produced by each of these methods for each graph order for degree centralization. As shown in the movie, as graph order increases, a bell-shaped pattern similar to that seen for closeness emerges regarding frequency of values obtained for degree. All of the methods demonstrate a peak at very low centralization levels (around 0.2). Unlike for closeness, Barabási-Albert preferential attachment models also have high ranks for centralization values between 0.8 and 1, but the number of bins with ranked values is small. Again, Star Start has moderate ranks for centralization values between 0.8 and 1 and the other methods all have very low ranks.

**Range of Graph Structures** Movie 3.9 further describes the distribution of degree centralization values by plotting them against the maximum degree centrality value for each method. As shown in the movie, two node type preferential attachment and Star Start have the greatest range of maximum centrality values for a particular centralization bin although both drop off significantly at high levels of centralization. The range for Barabási-

Albert preferential attachment models is very narrow. At low to moderate centralization levels, Erdös-Rènyi random graphs and small world graphs also have a broad range of maximum centrality values, although the small world graphs show some gaps between the maximum centrality values. Interestingly, there is a regular pattern between maximum centrality value and centralization value for small world graphs that is likely related to graph structure.

## 3.5 Discussion

Based on the results of these simulations, two node type preferential attachment and Star Start produce most of the full range of centralization values for all three measures of centralization that were studied. At high centralization levels for betweenness, Barabási-Albert preferential attachment, two node type preferential attachment, and Star Start produce a broad range of centralization values. However, Barabási-Albert preferential attachment produces fewer unique graphs with high centralization values for closeness, and even less well for degree. On the other hand, two node type preferential attachment and Star Start consistently produce a broad range of unique values at these high centralization levels, although it should be noted that performance does deteriorate as graph order approached 20 nodes.

Erdös-Rényi random graphs fail to produce highly centralized graphs as graph order increases beyond very small orders. This result is unsurprising given that highly centralized graphs are very rare in the full set of all possible random graphs. If nodes in a graph are labeled, then each graph can be uniquely identified. Then the sum of all possible graphs is:

$$\sum_{x=0}^{\frac{n(n-1)}{2}} \binom{\frac{n(n-1)}{2}}{x} = 2^{\frac{n(n-1)}{2}}.$$

Since the number of edges in a star graph is $n-1$, the number of graphs with $n-1$ edges is $\binom{\frac{n(n-1)}{2}}{n-1}$ with only $n$ of those graphs being star graphs (a star graph structure can be

produced $n$ times, once with each node as the center of the star). As the number of nodes increases, the fraction of graphs with $n-1$ edges in a star formation gets smaller.

Also unsurprising is the high rank of centralized graphs produced by the Barabási-Albert preferential attachment model. In this model, many nodes have few edges and a few nodes have lots of edges since the distribution of degree typically follows a power law. This means that although most Barabási-Albert preferential attachment graphs have one node that dominates, this generating method is much less likely to produce graphs with multiple nodes that have the maximum centrality thereby producing the full range of unique graphs that are highly centralized.

It is interesting to note that for all the graph generating methods, low centralization graphs are the most common. The structures that create highly centralized graphs, such as star graphs or star graphs with additional edges, have a low probability of occurring while there are many structures that produce low to moderate centralization.

### 3.5.1 Limitations

Due to computing limitations, this study evaluated graphs of small order (up to 20 nodes). However, the graph orders investigated correspond to small-group settings- an important environment for social network research. Additionally, due to the large numbers of graphs produced it was not feasible to compute the number of unique graphs produced by each method in a particular centralization bin. The number of unique maximum centrality values was used as a proxy for this, although clearly this is an underestimate. Also, this is a simulation study; although large numbers of graphs were generated it is possible that, if the study were repeated, slightly different results could be obtained. However, it seems unlikely that the conclusions would be be significantly different. Additionally, although there are a wide variety of graph generating methods, this study only focused on the most commonly utilized ones for analysis.

### 3.5.2 Future Directions

Since a method of generating graphs that span the range of centralization values for closeness, betweenness, and degree centralization has been found, future research can investigate the properties of centralization as graph order increases. At a descriptive level, the relationship between these three measures could be calculated as well as their relationship with the maximum centrality value. Since the calculation of centralization requires knowledge of the centrality for each node in the network, predicting centralization based on one or a few nodes could be useful for incompletely known networks.

## 3.6 Conclusion

Erdös-Rényi random graphs produce mostly low to moderately centralized graphs with a good range of graph structures. Graphs with high centralization values are unlikely to be generated by the algorithm. Similarly, the small world graph method produces graphs are only low to moderately centralized but does produce a range of graph structures. Barabási-Albert preferential attachment graphs can be highly centralized but do not produce a variety of graph structures with the same centralization value. Two node type preferential attachment and Star Start produce most of the full range of centralization values with a broad range of maximum centrality values. With the exception of the Barabási-Albert preferential attachment method, all of the graph generating methods examined produce the majority of graphs in the low to moderate centralization level, suggesting that those values are most easily generated and high centralization values are very uncommon.

## 3.7 Movies

### 3.7.1 Closeness

Movie 3.1. Closeness centralization values obtained for all graph generating methods (Erdös-Rènyi random graphs(ER), Barabási-Albert preferential attachment (BA), small world

(SW), two node type preferential attachment (Pref), and Star Start (Star)) for graphs of order 5-20 nodes.

(Loading Video...)

Movie 3.2. Rank of closeness centralization values obtained for all graph generating methods (Erdös-Rènyi random graphs (ER), Barabási-Albert preferential attachment (BA), small world (SW), two node type preferential attachment (Pref), and Star Start (Star)) for graphs of order 5-20 nodes.

(Loading Video...)

Movie 3.3. Scatter plot of maximum closeness centrality and closeness centralization values obtained for all graph generating methods (Erdös-Rènyi random graphs (ER), Barabási-Albert preferential attachment (BA), small world (SW), two node type preferential attachment (Pref), and Star Start (Star)) for graphs of order 5-20 nodes.

(Loading Video...)

### 3.7.2 Betweenness

Movie 3.4. Betweenness centralization values obtained for all graph methods (Erdös-Rènyi random graphs (ER), Barabási-Albert preferential attachment (BA), small world (SW), two node type preferential attachment (Pref), and Star Start (Star)) for graphs of order 5-20 nodes.

(Loading Video...)

Movie 3.5. Rank of betweenness centralization values obtained for all graph methods (Erdös-Rènyi random graphs (ER), Barabási-Albert preferential attachment (BA), small world (SW), two node type preferential attachment (Pref), and Star Start (Star)) for graphs of order 5-20 nodes.

(Loading Video...)

Movie 3.6. Scatter plot of maximum betweenness centrality and betweenness centralization values obtained for all graph methods (Erdös-Rènyi random graphs (ER), Barabási-Albert preferential attachment (BA), small world (SW), two node type preferential attachment (Pref), and Star Start (Star)) for graphs of order 5-20 nodes.

(Loading Video...)

### 3.7.3 Degree

Movie 3.7. Degree centralization values obtained for all graph methods (Erdös-Rènyi random graphs (ER), Barabási-Albert preferential attachment (BA), small world (SW), two node type preferential attachment (Pref), and Star Start (Star)) for graphs of order 5-20 nodes.

(Loading Video...)

Movie 3.8. Rank of degree centralization values obtained for all graph methods (Erdös-Rènyi random graphs (ER), Barabási-Albert preferential attachment (BA), small world (SW), two node type preferential attachment (Pref), and Star Start (Star)) for graphs of order 5-20 nodes.

(Loading Video...)

Movie 3.9. Scatter plot of maximum degree centrality and degree centralization values obtained for all graph methods (Erdös-Rènyi random graphs (ER), Barabási-Albert preferential attachment (BA), small world (SW), two node type preferential attachment (Pref), and Star Start (Star)) for graphs of order 5-20 nodes.

(Loading Video...)

## 3.8 Appendix

### 3.8.1 Movies

Movie 3.10. Closeness centralization values obtained for two node type preferential attachment (Pref) and Star Start (Star) methods for graphs of order 5-20 nodes.

(Loading Video...)

Movie 3.11. Betweenness centralization values obtained for two node type preferential attachment (Pref) and Star Start (Star) methods for graphs of order 5-20 nodes.

(Loading Video...)

Movie 3.12. Degree centralization values obtained for two node type preferential attachment (Pref) and Star Start (Star) methods for graphs of order 5-20 nodes.

(Loading Video...)

# Part III

# Topic 2

# Chapter 4

# Various Properties of Centralization in Small Graphs

## 4.1 Introduction

Centralization is a measure of how important one node is compared to all of the other nodes in a network. Centrality is a measure of how central/important a node is based on it's location in the graph. Unlike centrality, which is a node-level measure, centralization is a graph-level measure. Centralization relies on all of the node-level centrality information and as such is directly related to the overall structure of the graph. Previous work describes the centralization values produced by the common network generating methods and found that two node type preferential attachment and a new method called Star Start created graphs with the full range of centralization values for closeness, betweenness, and degree.[97] The goal of this paper is to further examine the properties of and the relationship between these centralization measures using graphs generating by the Star Start method.

Of all of the graph generating methods considered, only the $Gnm$ version of the Erdös-Rènyi random graph allows the calculation of the probability of obtaining labeled, undirected, and simple graphs with particular structures. To review, for a $Gnm$ Erdös-Rènyi random graph, a specified number of edges are randomly added to a network of a particular order. More

specifically, $m$ edges are drawn uniformly randomly from the set of all possible edges and added to the graph. In other words, $Gnm$ random graphs are chosen uniformly at random from the collection of all graphs which have $n$ nodes and $m$ edges. Note that the range of edges for a graph is 0, an empty graph, to $\frac{n(n-1)}{2}$, a complete graph. The sum of all possible graphs with (labeled edges) is:

$$\sum_{m=0}^{\frac{n(n-1)}{2}} \binom{\frac{n(n-1)}{2}}{m} = 2^{\frac{n(n-1)}{2}}$$

[98] Figure 4.1 describes the total number of Erdös-Rènyi $Gnm$ random graphs for graphs of order 5-8.

From information about the total number of graphs, it is possible to calculate the probability of obtaining a graph with a specified order and size (number of edges and nodes, respectively). Note that due to symmetry of a particular labeled graph structure, multiple graphs can have the same structure although with different node labels, called *graph isomorphism*. Isomorphic graphs have identical structure/topology but with different node labels. Clearly, if two graphs have the same structure then their node centrality and centralization values will be identical. The number of different *isomorphic classes*, or possible different symmetries, of a graph with a given order is determined by Pólya's Enumeration Theorem.[98]. Unfortunately, this enumeration does not provide information on the number of graphs within each class. Aspects of graph isomorphism such as counting all possible isomorphisms of one structure or determining if two graphs are isomorphic have been much studied but with few general results. In fact, determining the complexity of the graph isomorphism problem is even difficult.[99, 100]

The number of graphs with a particular structure is needed to precisely compute the probability of obtaining that graph. This tally would also be used to determine the probability of obtaining a centralization value based on that graph structure. For graphs with particularly well-studied structures, such as rooted graphs or directed graphs, it is possible to calculate the theoretical number of graphs within specific isomorphic classes.[98] Progress has been made calculating the number of isomorphisms of a particular graph using various computer

programs, such as *nauty*.[101] Thus, either theoretical or computational methods only allow the enumeration of the number of graphs with one graph structure. Unfortunately, neither of these methods allows the direct listing of all possible Erdös-Rènyi $Gnm$ random graphs of a particular order. As a result, there is no simple solution for directly finding the complete distribution of graph structures for a given graph order, either computationally or theoretically, although you can sample from it.

In theory, the set of all $Gnm$ Erdös-Rènyi random graphs for a particular graph order provides the true distribution of graph structures in random graphs of that order since the probability of obtaining any one structure can be calculated. In this way, the same set of all $Gnm$ Erdös-Rènyi random graphs of a particular graph order also provides the true distribution of centralization values for random graphs. During a discussion of graph structures, it is important to note that widely different graph structures can produce the same centralization value. For example, a centralization value of 0 is obtained by any graph where all of the nodes are equivalent, such as empty, ring, or fully connected graphs. Excluding this example, the exact relationship mapping all possible graph structures to their particular centralization value has not been investigated. Given the difficulty in determining the complete distribution of graph structures and the unknown relationship between number of different structures that produce the same centralization value, this study will use a sampling method to determine the distribution of centralization values for Erdös-Rènyi $Gnm$ random graphs

When considering centralization, a star graph has the maximum possible centralization value for closeness, betweenness, and degree centralization (see Measures section below). Using the above formula, the probability of randomly obtaining star graph can be computed as: $\frac{n}{2^{\frac{n(n-1)}{2}}}$. The numerator is $n$ and not one because a star graph has $n$ isomorphisms since each of the $n$ nodes could be the center node of the star. Note that rotating the peripheral nodes does not produce an isomorphic graph because the adjacency matrix is the same for both the rotated graph and the original graph. Clearly, as the number of nodes increases even minimally, the fraction of graphs with $n-1$ edges in a star formation goes to zero fairly rapidly. For example, the probability of randomly obtaining a five node star graph

is 0.004882812 and decreases substantially to $2.842171e^{-13}$ for a 10 node star graph. Even if attention is focused on producing random graphs with $n-1$ edges instead of the entire set of random graphs for a particular graph order, the number of graphs with $n-1$ edges is $\binom{\frac{n(n-1)}{2}}{n-1}$ with only $n$ of those graphs being star graphs. This fraction also goes to zero as graph order increases to even moderate levels. For example, the probability of obtaining a five node star graph in the set of graphs with $n-1$ edges is 0.02380952 and decreases to $1.12846e^{-8}$ for a 10 node star graph.

In order to make any inferences about closeness, betweenness, and degree centralization and their various properties, the true distribution of centralization values must be obtained. As discussed earlier, no simple solution exists for finding the exact distribution of graph structures, and therefore the exact distribution of centralization values. As such, this study employs a sampling method to try to ensure that the full range of centralization values is obtained. However, as illustrated above with the example of the star graph, some graph forms are very rare and have a very low probability of being randomly obtained. One way around this is to produce a large number of graphs in order to increase the probability of obtaining these rare graph structures. The probability of randomly obtaining a non-star graph when the number of edges is set to $n-1$ is $1-\frac{n}{\binom{\frac{n(n-1)}{2}}{n-1}}$. Thus, when the number of iterations is increased from one to $x$, the probability of obtaining a star graph is $1-\left(1-\frac{n}{\binom{\frac{n(n-1)}{2}}{n-1}}\right)^{x}$. This means that an extremely large number of random graphs must be generated to increase the probability of randomly producing graphs of even moderate orders with highly centralized forms. For example, the probability of randomly obtaining a 10 node star graph in one interaction is $1.12846e^{-8}$. In order to have this probability approach one, over 600 million graphs are required. See Figure 4.2 for the probability of a star graph for smaller graph orders. As a result, a Monte Carlo approach to generating the distribution of centralization values using Erdös-Rènyi *Gnm* random graphs is not feasible since it would require generating a very, very large number of graphs even if the number

of edges is set. This is highly inefficient and not conducive to a study that examines the properties of centralization. An alternative method is to use the previously described Star Start method as a means of creating a pseudo-random sample of possible graph forms. The idea behind Star Start is that each iteration from star graph to complete or empty graph is a pseudo-random sample of the full range of centralization values that are possible (pseudo-random because the end-points are fixed). Importantly, this sample reflects the underlying distribution of centralization values because the graph forms are randomly obtained by adding or deleting edges. Then the sample average across the iterations for various statistics can be computed. For this analysis the starting graph, the star graph, is dropped since it artificially inflates the number of graphs with the highest possible centralization value.

### 4.1.1 Measures

For this paper we will call a particular network a *graph*, *nodes* are the points in the graph and *edges* connect the nodes. The number of edges in a graph is called the graph's *size* while the number of nodes in a graph is called the graph's *order*. A *path* is the sequence of edges between two nodes. The *geodesic* is the shortest path between two nodes, where distance is defined as the count of number of edges. A graph is *connected* if there is a path between each node in the graph and *disconnected* otherwise. Graph *isomorphisms* refer to graphs that have the same structure although the edges creating the graph are different. A graph is *simple* if there is only one edge connecting nodes (multiple or weighted edges are not allowed). A *digraph* is a directed graph, where the edges are directed from one node to another. Otherwise, edges are *undirected*. *Self-loops* are edges that start and end with the same node.

This paper uses the relative centrality and centralization measures for closeness, betweenness, and degree to facilitate comparisons between graphs of different orders. All of these measures were described by Freeman in his much referenced 1978 paper.[6] Relative measures are obtained by taking the observed value and dividing by the theoretical maximum value for that measure. Using relative measures, the range of possible values falls between zero and one. For closeness, betweenness, and degree centrality, the theoretical maximum

has been proven to be achieved by the center node of a star graph.[6] For closeness, betweenness, and degree centralization the theoretical maximum has been proven to be achieved by the star graph.[6, 8] Although the highest possible relative centralization is given by a star graph, centralization values $\geq 0.8$ can be obtained by adding a small number of edges to a star graph such that there are multiple nodes with the highest possible centrality and many more nodes with the lowest possible centrality.

**Closeness** Let $d(n_i, n_k)$ be the number of edges in the geodesic between node $n_i$ and node $n_j$. By convention, if node $n_i$ and node $n_j$ are not connected by any edges, then $d(n_i, n_j) = \infty$. Of course, the number of edges between node $n_i$ and itself is 0, so $d(n_i, n_i) = 0$. Then the relative closeness centrality (called *closeness centrality* from here forward) of node $n_i$(in an unweighted connected graph) is defined as $C'_c(n_i) = \left[\frac{\sum_{j=1}^{n} d(n_i,n_j)}{n-1}\right]^{-1} = \frac{n-1}{\sum_{j=1}^{n} d(n_i,n_j)}$.[6] In his 1978 paper, Freeman shows that the maximum closeness centrality measure is obtained by the central node in a star graph. By definition, closeness centrality can only be computed for a connected graph (where each node is connected by at least one edge). However, *igraph* substitutes $n$ for $\infty$ in the case of disconnected nodes.[10, 11] Consequently, closeness centrality can be computed for all graphs, connected or disconnected.

Relative closeness centralization (called *closeness centralization* from here forward) is: $C_c = \frac{\sum_{i=1}^{n}[C'_c(n^*)-C'_c(n_i)]}{(n^2-3n+2)/(2n-3)}$, where $C'_c(n^*)$ is the maximum closeness centrality of the observed graph. Given this formula, it is clear that a centralization of 0 is obtained by a graph where all nodes are equal (such as a ring graph, an empty graph, or a complete graph).

**Betweenness** Any nodes on the geodesic connecting two nodes are said to be between them. Betweenness centrality measures how often a node is between other nodes. $C_B(n_k) = \sum_{i}^{n}\sum_{<j}^{n} b_{ij}(n_k)$ where $b_{ij}(n_k) = \frac{g_{ij}(n_k)}{g_{ij}}$ is the number of geodesic connecting $n_i$and $n_j$ that contain $n_k$ and $g_{ij}$ is the total number of geodesics connecting $n_i$ and $n_j$. Using the center point in a star graph as the reference, relative betweenness centrality (*betweenness centrality*) is defined as:$C'_B(n_k) = \frac{2*C_B(n_k)}{n^2-3n+2}$.

Relative betweenness centralization (*betweenness centralization*) is calculated as: $C_B =$

$\frac{\sum_{i=1}^{n}[C'_B(n^*)-C'_B(n_i)]}{n-1}$ where $C'_B(n^*)$ is the maximum betweenness centrality of the observed graph.

**Degree** The number of edges connected to a node is the degree of the node. Degree centrality is defined as: $C_D(n_k) = \sum_{i=1}^{n} a(n_i, n_k)$ and relative degree centrality (called *degree centrality* from here forward)is defined as:
$C'_D(n_k) = \frac{\sum_{i=1}^{n} a(n_i,n_k)}{n-1}$ where

$$a(n_i, n_k) = \begin{cases} 1, & \text{iff edge between } n_i \text{ and } n_k \\ 0, & \text{otherwise} \end{cases}$$

Relative degree centralization (*degree centralization*) is calculated as:
$C_D = \frac{\sum_{i=1}^{n}[C_D(n^*)-C_D(n_i)]}{n^2-3n+2}$ where $C_D(n^*)$ is the maximum degree centrality of the observed graph.

## 4.2 Methods

### 4.2.1 Graph Generating Methods

**Erdös-Rènyi Random Graphs** In the *Gnm* version of the Erdös-Rènyi random graph, a specified number of edges are randomly added to a network of a particular order. More specifically, $m$ edges are drawn uniformly randomly from the set of all possible edges and added to the graph. Since the goal of this research is to determine the range of centralization values possible for a particular network generating method, centralization was measured as edges were incrementally added. The number of edges of the network ran from $|E(g)| = 0$ (creating an empty graph with no edges linking any nodes) to $|E(g)| = \frac{n(n-1)}{2}$ (creating a complete graph with the maximum possible number of edges), incrementing by one edge. Graphs were undirected and multiple edges between two nodes and self-loops were not allowed. Unfortunately, as noted above in the Introduction, there is no simple way to find

the distribution of $Gnm$ Erdös-Rènyi random graphs for a particular order. Consequently, this paper will utilize a method that generates equal numbers of $Gnm$ Erdös-Rènyi random graphs for each edge set of a given size. For each graph order, the number of repetitions (randomly obtained graphs) required to obtain a star graph with probability 0.999 when the number of edges was set to $n-1$ was computed and used as the number of repetitions for that graph order. All three centrality and centralization measures were calculated for each graph produced. Note that $Gnm$ graphs are not required to be connected. Due to computing restrictions, only graphs with 5-7 nodes were constructed.

**Star Start** The method starts with a star graph, $g$, with $n$ nodes and then edges were randomly added or deleted (with equal probability) from the graph until $g$ was complete or empty. Thus, in order to complete an iteration, either a minimum of $n-1$ edges must be removed from the star graph to create an empty graph or a minimum of $(n-1)(n-2)/2$ edges must be added to the star graph to create a complete graph. Loops and duplicate edges were not allowed. All three measures of centrality and centralization were taken for each update of $g$. This process was repeated 500 times (iterations), each time starting with a star graph with $n$ vertices and continuing until the graph was complete or empty.

### 4.2.2 Comparison of Distributions

In order to compare graphs of different orders, centralization values were binned into 101 bins from 0 to 1 in increments of 0.01. Since the first graph of the Star Start method is fixed as a star graph, it is dropped from the analysis for all iterations. Considering each Star Start iteration as a sample of the possible graph centralization values, the average number of graphs in a centralization bin is used for all calculations. A Kolmogorov-Smirnov test was used to compare the distribution of centralization values for the Erdös-Rènyi and Star Start methods of generating graphs for each of closeness, betweenness, and degree centralization. Visual inspection of the distributions was conducted by plotting a contour plot of the centralization values side-by-side.

### 4.2.3 Centralization Associations

As in the comparison of distributions, centralization values were binned into 101 bins from 0 to 1 in increments of 0.01. Since the first graph of the Star Start method is fixed as a star graph, it is dropped from the analysis for all iterations. A histogram was used to describe the distribution of centralization values for closeness, betweenness, and degree centralization. For the histogram, the proportion of graphs that fell into each centralization bin was computed for each iteration (proportion and not absolute number was used to control for the variable number of graphs in each iteration). Then the average of these 500 values was taken for each centralization bin and plotted in the histogram. A three-dimensional scatter plot was used to visualize the closeness, betweenness, and degree centralization values for each graph obtained for this study. To ease interpretation of the three dimensional plots, the coloring of the points in the plot is done such that lighter colors are closer to the front of the graph and darker colors are farther away. To better understand the graph structures that produced the shape of the scatter plot, centralization values were dichotomized into two groups: centralization $< 0.4$ and centralization $\geq 0.4$ for closeness, betweenness, and degree centralization. Note that the 0.4 cut-off represents moderate centralization and is arbitrarily chosen for illustrative purposes. The scatter plot was re-colored to emphasize placement of graphs into specific three-way categories. Pearson product-moment correlation was used to calculate linear association between two measures for each of the 500 iterations and then the average taken.

Linear regression with both one and two of the centralization measures as predictors and another as the dependent variable were also performed for each centralization measure and each of the 500 iterations. The average coefficients and $R^2$ value were taken across all of the iterations. Average $R^2$ was used to compare models of the same order. For the models with only one predictor, the regression line was plotted against the minimum, mean, and maximum centralization values. For the multiple linear models, the regression plane was plotted against the centralization values. For both plots the opacity of the points reflects the number of graphs that fell into a particular centralization bin. Darker points have larger numbers of graphs and lighter points have fewer numbers of graphs. Additional linear

models considered average centrality and maximum centrality as predictors of centralization. The same methods were used as described above for centralization as the predictor.

Analyses were performed for each graph order between 5 and 20 nodes and results summarized by graph order. All computations were performed in R version 2.15.1 with the *igraph* package version 0.6.2.[73, 72] Significance was set at $\alpha < 0.05$ and all tests were two-sided.

## 4.3 Results

### 4.3.1 Comparison of Distributions

Figure 4.3 describes the results of the Kolmogorov-Smirnov tests comparing the distribution of centralization values between Erdös-Rènyi and Star Start methods of generating graphs for closeness, betweenness, and degree centralization. There is no significant difference between the distribution of the centralization values for the two methods. Figure 4.4 compares the actual distribution of centralization values for Erdös-Rènyi and Star Start graphs for graph of order 5 to 7 nodes for closeness centralization using a contour plot. The colors of the plot (pink to blue) reflect a fewer relative numbers of graphs that fell into a particular centralization bin. Figures 4.5 and 4.6 illustrate the same comparison for betweenness and degree centralization, respectively.

### 4.3.2 Centralization Associations

Figures 4.7, 4.8, and 4.9 describe the distribution of centralization values for closeness, betweenness, and degree using a contour plot. Figure 4.10 provides the $5th$ percentile and $95th$ percentile of the distribution for each of the measures by each graph order analyzed.

Movie 4.1 illustrates the scatter plot of closeness, betweenness, and degree centralization for each graph order. As shown in the Movie, as graph order increases the scatter plot has several noticeable projections or "fingers". Movie 4.2 provides a closer examination of

graph structure in the fingers of the scatter plot for graphs of order 5-8 nodes. As shown in the movie, each finger on the left-side of the plot contains graphs with a specific number of disconnected nodes. The main finger on the right-side of the scatter plot contains graphs with high closeness, betweenness, and degree centralization. The base of the main finger predominantly contains graphs where either betweenness is low and closeness and degree centralization are high or degree is low and closeness and betweenness are high. There are also a few connected graphs where closeness is low but betweenness and degree centralization are high.

In order to more clearly see the relationship between each pair of centralization measures, Movies 4.14-4.16 in the Appendix provide the scatter plot of values for each pairwise combination for each graph order as well as the corresponding Pearson correlation coefficient. Linear and quadratic models using one centralization measure as the dependent variable and another as the independent variable did not fit the data well and are not reported in this paper. A better, but still poor fit for most of the models, is a linear model fit to the dependent centralization measure restricted to be $\geq 0.6$. Movies 4.3 and 4.4 describe the two models that fit reasonably well using the restricted centralization measure as the outcome (the other more poorly fitting models are not shown). Additional models with one centralization measure as the dependent variable and the remaining two as the independent variables were investigated and are shown in Movies 4.5-4.7.

Next, a linear model with maximum centrality as the predictor and the corresponding centralization measure as the outcome was considered. Linear, quadratic, and cubic terms for maximum centrality were evaluated. The results were similar but the model with quadratic maximum centrality fit slightly better than the others. Movies 4.8-4.10 describe the minimum, mean, and maximum maximum centrality value for each centralization bin and the regression line for each graph order. Movies 4.17-4.19 in the Appendix illustrate the scatter plot of maximum centrality and centralization with the corresponding Pearson correlation coefficient for graphs of each order for closeness, betweenness, and degree. As before, opacity of the points reflects the number of graphs that fell into a particular centralization bin.

Finally, a linear model with average centrality as the predictor and the corresponding centralization measure as the outcome was considered. Linear, quadratic, and cubic terms for maximum centrality were evaluated. As with the model using maximum centrality as the predictor, the results were similar but the model with quadratic maximum centrality fit slightly better than the others. Movies 4.11-4.13 describe the minimum, mean, and maximum average centrality value for each centralization bin and the regression line for each graph order. Movies 4.20-4.22 in the Appendix illustrate the scatter plot of average centrality and centralization with the corresponding Pearson correlation coefficient for graphs of each order for closeness, betweenness, and degree. As before, opacity of the points reflects the number of graphs that fell into a particular centralization bin.

## 4.4 Discussion

The graphs produced by the Start Start method, except the starting graph, can be considered a pseudo-random sample of the range of centralization values. Analyses suggest that the Star Start method provides a relatively more efficient mechanism of generating highly centralized graphs than the ER method while still following the same distribution of centralization values for closeness and degree. Betweenness results suggest some differences between the distributions but these could possibly be addressed by modifications to the Star Start program or by running more iterations of the program. Admittedly, Star Start and ER have different goals; the Star Start program generates graphs to produce a range of centralization values while ER method generates graphs with uniform probability for each (labeled) structure. As such, the Star Start program can only approximate the distribution of centralization values, not the distribution of graph structures.

The Star Start program can be considered to generate a null distribution for comparing centralization values from real-life networks.As shown in Figure 4.10 and in Figures 4.7, 4.8, and 4.9 graphs with high centralization values, centralization $> 0.80$, are rare at all graph orders. As graph order increases, the $95th$ percentile continues to decrease until graphs with centralization values $> 0.6$, a moderate cutoff value, are very rare. This downward trend

suggests that for larger graph orders than those studied here, having at least a moderate centralization value is unusual.

Unfortunately, for validation purposes, only Star Start graphs with very small order could be compared to Erdös-Rènyi *Gnm* random graphs due to computing limitations. However, the lack of apparent differences in the distributions as order increased slightly suggests that the findings should be true for the orders considered in this analysis. The Star Start method appears to work less well for betweenness centralization as network order increases, although no significant differences were found using the Kolmogorov-Smirnov test. The problem appears to be with higher than expected numbers of graphs with moderate to high centralization levels. Examination of the histogram for closeness and degree also show an increased number of graphs in high centralization bins, although to a lesser extent. It is unclear why this problem only appears for betweenness and not for closeness or degree centralization, although the distribution for betweenness centralization is more skewed towards zero than the others. As a possible explanation for these findings, the number of graphs generated for the Erdös-Rènyi *Gnm* graphs as the true distribution of centralization values was designed to produce at least one star graph with high probability, and not necessarily to produce all of the possible graph structures with high probability. Thus, it is possible that highly centralized graphs (centralization between 0.8 and 1) were not obtained at a level representative of their probability. Alternatively, the Star Start program may overemphasize higher centralization graphs since by design all graphs move from a centralization of 1 to 0.

It should also be noted that the average of the Star Start program could only be compared to a sample of Erdös-Rènyi *Gnm* random graphs rather than the entire distribution. As discussed in the Introduction, there is no simple way to produce the complete Erdös-Rènyi *Gnm* random graph distribution for a graph of a particular order. However, the sampling method utilized in this paper ensured that the full range of centralization values was obtained. A sampling scheme that reflected the proportional number of graphs for each graph order (rather than a uniform number of graphs generated for each graph order) was also considered and the distributional results were similar.

Even though the average of the Star Star iterations may over-emphasize higher centralized graphs, as shown in Movie 4.1, all combinations of centralization $\geq 0.6$ are obtained. Interestingly, there are noticeable gaps in combinations above this level producing fingers in the scatter plot. Also, as graph order increases, there are a broad range of graphs with low closeness but high degree and betweenness centralization and fewer graphs with high values for all three centralization measures. This can be seen on the scatter plot as the main body of points splits into several different fingers with one main fork for graphs with low closeness centralization but high degree and betweenness centralization and another fork for graphs with high values for all three centralization measures. The fingers on the left-side of the plot (those graphs with low closeness but high betweenness and degree centralization) are disconnected graphs. The important feature of these scatter plots is that each finger contains a specific number of disconnected nodes. The placement of these fingers away from the main fork of the graph on the right-hand side of the plot is likely due to the imputation method used for calculating closeness centrality in disconnected graphs. An interesting study would be to see if different imputation methods for closeness centrality in disconnected graphs affect the placement of the fingers.

For all three pairwise comparisons of centralization measures, the correlation decreases from a strong correlation to a moderate correlation as graph order increases. Although a Pearson correlation is reported, a Spearman correlation also produced similar results. These results are consistent with the scatter plots which demonstrates a very broad range of values at low to moderate centralization levels as graph order increases. The more narrow range at higher centralization levels led to the restricted model, but most of the models attempted did not fit the data very well. However, the models predicting restricted degree centralization based on betweenness centralization and restricted closeness centralization predicted by degree centralization fit the data reasonably well. Furthermore, the regression equation for predicting restricted closeness centralization by degree centralization did not change very much as the order of the graph increased suggesting that this relationship might hold for larger orders as well. Unfortunately, the regression equation for predicting restricted degree centralization by betweenness centralization changed as graph order increased, although

some extrapolations might be made for larger graph orders since the changes followed a general trend.

The models with two centralization measures as predictors also fit reasonably well, although $R^2$ decreased as order increased. Interestingly, $R^2$ decreased or stayed the same for all but two of the models with one centralization measure predictor. In those two models, both involving betweenness and degree centralization, it increased as order increased.

The models with maximum centrality as the predictor fit fairly well, especially when the maximum centrality values was $\leq 0.6$. For all three centralization measures, a quadratic model fit best. The coefficients for all of the models stayed relatively constant suggesting that these models may be true for larger graph orders as well. The model for betweenness centralization fit particularly well, likely due to the much more restricted range of centralization values corresponding to a particular maximum centrality value. For maximum centrality values $> 0.6$, the models overestimated the mean centralization value for closeness and degree.

Surprisingly, the models with average centrality as the predictor do not fit as well as models based on maximum centrality even though they are based on more information. Again, for all three centralization measures, a quadratic model fit best, although none of the models fit particularly well, especially after average centrality increased past the average centrality of a star graph.

**Uses** The main goal for these prediction models is provide a simple model for applied network researchers to calculate centralization given some limited information about their network. Centralization is not calculated by many of the network software programs commonly used by applied network researchers, such as Gephi and NodeXL.[74, 75] Models that predict centralization based on the information reported by these programs, such as the one maximum centrality value or average centrality, will be useful for applied network researchers utilizing programs that only provide limited centrality information. With this in mind, more complicated statistical models were not attempted as the goal is provide a reasonable estimate of centralization with minimal information.

**Limitations** Since it is not computationally feasible to generate the exact Erdös-Rènyi $Gnm$ random graph distribution for a particular graph order, this study uses a sampling method whereby equal numbers of graphs are generated for each edge set of a given size. However, as noted in the introduction, the number of graphs possible for a given edge size, $m$, ($m \in [0, \frac{n(n-1)}{2}]$), is $\binom{\frac{n(n-1)}{2}}{m}$ which is clearly not uniform. However, since centralization is directly related to graph structure, not number of edges, it is not clear that this sampling scheme is biased.

This is a simulation study and as such results may change if the Erdös-Rènyi $Gnm$ and Star Start programs were run again to generate new graphs for analysis. A preliminary investigation into this matter for graphs with 5 nodes has shown that the numbers of graphs generated in a particular centralization vary only slightly (less than 10% between different runs of the programs). Consequently, centralization bins from graphs that are very rarely obtained (number of graphs $< 20$) may or may not be present for an analysis. However, since very large numbers of graphs are generated by both programs it is unlikely that the set of new graphs would change the findings significantly. It should be noted that five hundred iterations were chosen for the Star Start program as generating enough graphs to approximate Erdös-Rènyi $Gnm$ random graphs while still being computationally feasible for the larger graph orders investigated. A preliminary investigation has shown that the proportion of graphs that fall into a particular centralization bin is different by 0.01 even after 500 iterations for graphs of order 5-13 nodes.

Also, due to computing limitations the results are for small graph orders. However, centralization as a measure for public health networks seems most applicable to small group settings. Lastly, the association and prediction models were generated based on the distribution of centralization in random graphs, not the true distribution of centralization values in real-life networks. Unfortunately, the true distribution of centralization values in real-life networks is unknown so the best alternative is to evaluate centralization in networks whose properties are known.

**Future Directions** Future areas of research include investigating these relationships in larger graph orders, considering other centrality/centralization measures, and prediction models based on slightly more information (such as the top $x$ percent of nodes). Of particular interest is simulating infections or information transfer across graphs with a range of centralizations. The results of such a study could be useful for public interventions applied to small group settings.

## 4.5 Conclusion

The average distribution of centralization values across 500 iterations of the Star Start program approximates the true distribution of centralization. As a result, some inference about centralization properties can be drawn using the Star Start data. Since the correlation decreases as graph order increases, it is unlikely that knowledge of one centralization measure will help researchers predict what a different centralization measure might be for the same network except in two very specific circumstances. Unfortunately, this means that each measure of centralization would need to be calculated on a network. Models predicting centralization based on maximum centrality perform reasonably well, especially when the maximum centrality value is $\leq 0.6$. Models based on average centrality fit poorly after the average increases past the average centrality of a star graph.

## 4.6 Figures, Movies

| Graph Order | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| # ER Gnm graphs | 1,024 | 32,768 | 2,097,152 | 268,435,456 |

Figure 4.1: Total number of Erdös-Rènyi $Gnm$ random graphs by graph order

| Order of graph (# of nodes) | Overall star graph probability | Star graph probability when edges = n-1 | # graphs required to produce star with 0.999 prob |
|---|---|---|---|
| 5 | 0.0048828 | 0.0238095 | 500 |
| 6 | 0.0001831 | 0.001998 | 4,000 |
| 7 | 3.34E-06 | 0.000129 | 54,000 |
| 8 | 2.98E-08 | 6.76E-06 | 1,050,000 |
| 9 | 1.31E-10 | 2.97E-07 | 25,000,000 |

Figure 4.2: Probability of randomly obtaining a star graph

### 4.6.1 Comparison of Distributions

| Order of graph (# of nodes) | KS p-value | | |
|---|---|---|---|
| | Closeness | Betweenness | Degree |
| 5 | 0.6994 | 0.6208 | 0.8690 |
| 6 | 0.8839 | 0.4506 | 0.8326 |
| 7 | 0.6164 | 0.0666 | 0.2221 |

Figure 4.3: Kolmogorov-Smirnov tests for distribution of Erdös-Rènyi random graphs and Star Start graphs

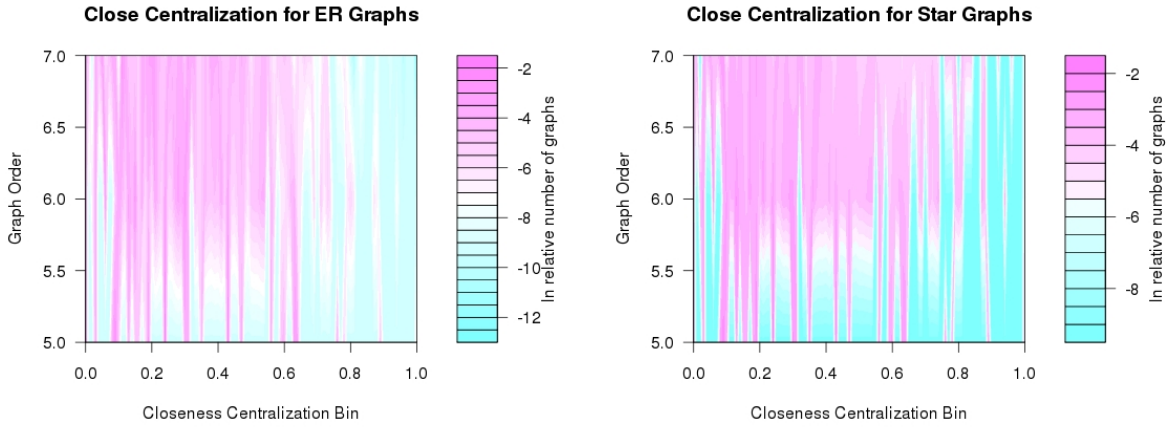


Figure 4.4: Contour plot illustrating the distribution of closeness centralization values for Erdös-Rènyi and Star Start graphs for graph of order 5 to 7 nodes.

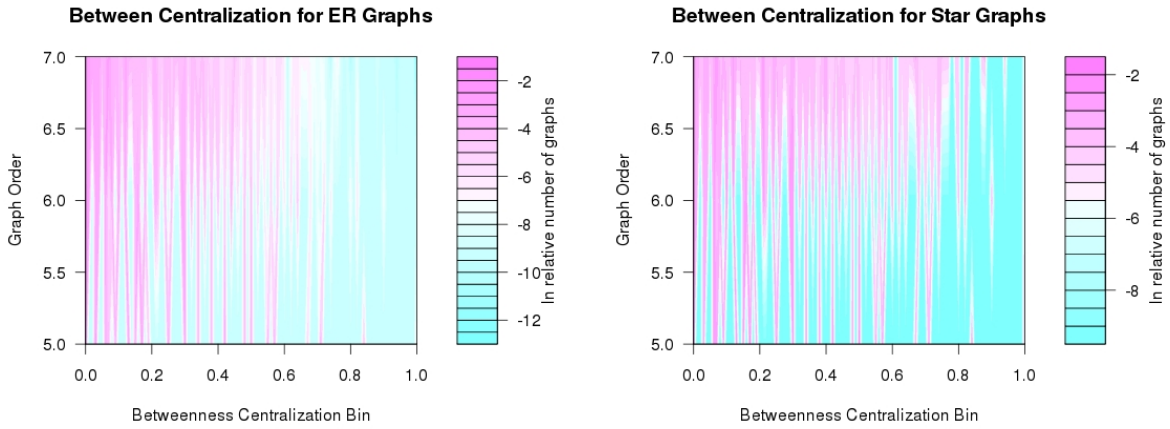


Figure 4.5: Contour plot illustrating the distribution of betweenness centralization values for Erdös-Rènyi and Star Start graphs for graph of order 5 to 7 nodes.

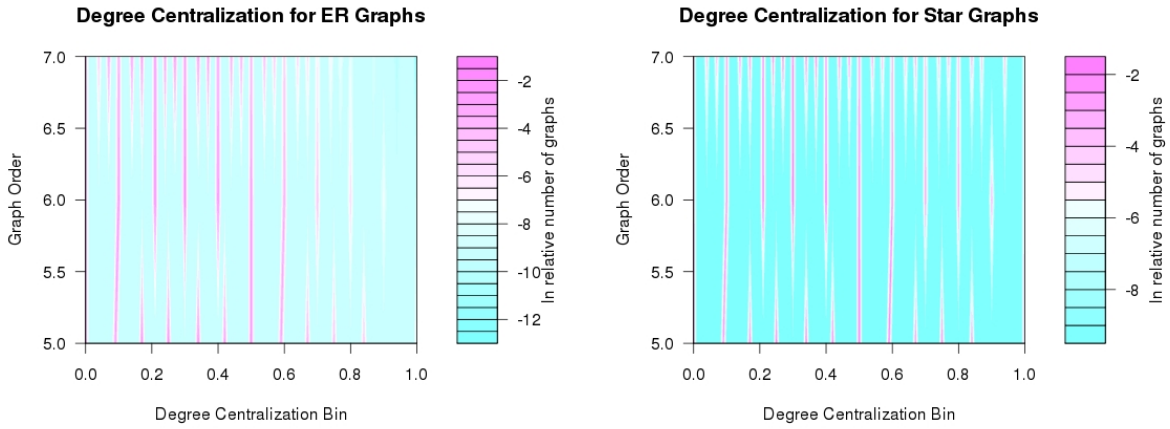


Figure 4.6: Contour plot illustrating the distribution of degree centralization values for Erdös-Rènyi and Star Start graphs for graph of order 5 to 7 nodes.

### 4.6.2 Centralization Associations

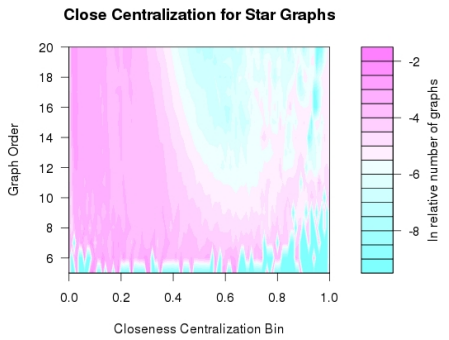


Figure 4.7: Distribution of closeness centralization values for graph of order 5 to 20 nodes.

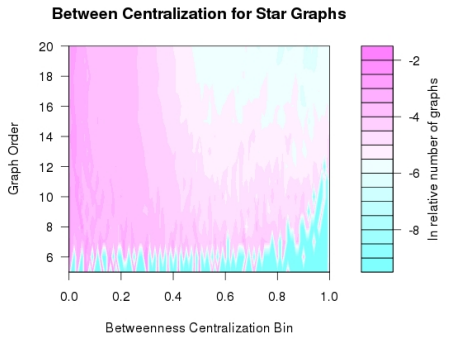


Figure 4.8: Distribution of betweenness centralization values for graph of order 5 to 20 nodes.

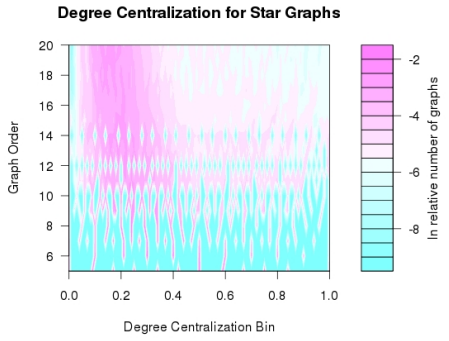


Figure 4.9: Distribution of degree centralization values for graph of order 5 to 20 nodes.

| Order of graph (# of nodes) | Centralization | | | | | |
|---|---|---|---|---|---|---|
| | Betweenness | | Closeness | | Degree | |
| | 5%tile | 95%tile | 5%tile | 95%tile | 5%tile | 95%tile |
| 5 | 0.00 | 0.84 | 0.00 | 0.89 | 0.00 | 0.84 |
| 6 | 0.00 | 0.74 | 0.03 | 0.79 | 0.10 | 0.80 |
| 7 | 0.00 | 0.67 | 0.03 | 0.68 | 0.10 | 0.77 |
| 8 | 0.00 | 0.72 | 0.03 | 0.65 | 0.10 | 0.77 |
| 9 | 0.00 | 0.68 | 0.03 | 0.58 | 0.09 | 0.74 |
| 10 | 0.00 | 0.68 | 0.02 | 0.53 | 0.09 | 0.73 |
| 11 | 0.00 | 0.66 | 0.02 | 0.49 | 0.09 | 0.72 |
| 12 | 0.00 | 0.64 | 0.02 | 0.45 | 0.10 | 0.70 |
| 13 | 0.01 | 0.63 | 0.02 | 0.43 | 0.09 | 0.68 |
| 14 | 0.01 | 0.65 | 0.02 | 0.40 | 0.08 | 0.70 |
| 15 | 0.01 | 0.60 | 0.02 | 0.38 | 0.08 | 0.65 |
| 16 | 0.01 | 0.59 | 0.02 | 0.37 | 0.08 | 0.63 |
| 17 | 0.01 | 0.58 | 0.02 | 0.36 | 0.08 | 0.62 |
| 18 | 0.01 | 0.56 | 0.02 | 0.35 | 0.08 | 0.58 |
| 19 | 0.01 | 0.57 | 0.01 | 0.34 | 0.08 | 0.58 |
| 20 | 0.01 | 0.57 | 0.01 | 0.33 | 0.08 | 0.58 |

Figure 4.10: $5th$ and $95th$ percentiles of centralization distributions for graphs of order 5 to 20 nodes

Movie 4.1. Scatter plot of closeness, betweenness, and degree centralization values obtained for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.2. Examination of fingers in scatter plot of closeness, betweenness, and degree centralization values obtained for graphs of order 5-8 nodes.

(Loading Video...)

Movie 4.3. Regression line for restricted closeness centralization predicted by degree centralization values for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.4. Regression line for restricted degree centralization predicted by betweenness centralization values for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.5. Regression plane for closeness centralization predicted by betweenness and degree centralization values for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.6. Regression plane for betweenness centralization predicted by closeness and degree centralization values for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.7. Regression plane for degree centralization predicted by closeness and betweenness centralization values for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.8. Regression line for closeness centralization predicted by maximum closeness centrality for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.9. Regression line for betweenness centralization predicted by maximum betweenness centrality for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.10. Regression line for degree centralization predicted by maximum degree centralization for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.11. Regression line for closeness centralization predicted by average closeness centrality for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.12. Regression line for betweenness centralization predicted by average betweenness centrality for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.13. Regression line for degree centralization predicted by average degree centralization for graphs of order 5-20 nodes.

(Loading Video...)

## 4.7 Appendix

### 4.7.1 Movies

Movie 4.14. Scatter plot of closeness and betweenness centralization values obtained for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.15. Scatter plot of closeness and degree centralization values obtained for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.16. Scatter plot of betweenness and degree centralization values obtained for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.17. Scatter plot of maximum closeness centrality and closeness centralization values obtained for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.18. Scatter plot of maximum betweenness centrality and betweenness centralization values obtained for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.19. Scatter plot of maximum degree centrality and degree centralization values obtained for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.20. Scatter plot of average closeness centrality and closeness centralization values obtained for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.21. Scatter plot of average betweenness centrality and betweenness centralization values obtained for graphs of order 5-20 nodes.

(Loading Video...)

Movie 4.22. Scatter plot of average degree centrality and degree centralization values obtained for graphs of order 5-20 nodes.

(Loading Video...)

# Part IV

# Topic 3

# Chapter 5

# Centralization in Disease Spread

## 5.1 Introduction

A common model for the spread of an infectious disease is the $S-I-R$ compartment model of disease transmission. In this model, individuals move from *Susceptible (S)* category to *Infected (I)* category to *Recovered/Removed (R)* category over time based on probabilities of infection transmission and recovery. Categories are mutually exclusive and exhaustive. An individual who is *susceptible* does not currently have the disease and is also not immune to becoming infected. An individual who is *infected* currently has the disease and can spread it to any susceptibles. An individual who is *recovered* has already had the infection and is currently immune to the disease and no longer able to spread the infection. The outcome of the $SIR$ model at each time point is the number of individuals in each category.

In order to implement a $SIR$ model, several starting values are needed: total population size, $N$, initial number of susceptible individuals, $S$, initial number of infectious individuals, $I$, and initial number of recovered individuals, $R$ (where $N = S + I + R$). Since individuals move from infected to recovered with no possibility of becoming reinfected, the basic $SIR$ model is best suited for diseases that confer immunity (or death) after infection. Thus, $SIR$ models are most appropriate for influenza-type illnesses or childhood diseases like measles. Of course, compartment models of infectious disease transmission can

vary according the characteristics of the infectious disease being modeled. Thus, there are $SIS$ models (Susceptible-Infectious-Susceptible) and $SLIR$ (Susceptible-Latent-Infectious-Recovered) models and others as well. $SIS$ models are appropriate for diseases where no immunity from future infection is gained from current infection, e.g. the common cold.

Common parameters of compartment models include $\beta$, the rate at which susceptible individuals become infected (also called the infection rate), $\kappa$ the rate of contact, and $\gamma$, the rate of recovery from infection.[102, 103, 104] If the population is closed, or static, then the total population size does not change and compartment change is strictly based on $\beta$, $\kappa$, and $\gamma$. However, if the population is open, or dynamic, then the total population size is allowed to change over time aside from the epidemic parameters. Births, deaths, immigration, and emigration can change the total number of individuals over time. Note that the rate of infection and recovery are properties of the infectious disease but can vary due to population behaviors.[1, 104]

Much research on disease spread in populations is based on a set of differential mean-field, or mass-action, equations that represent the movement between disease compartments over time:

$$\begin{aligned}
\frac{ds}{dt} &= -\beta\kappa si \\
\frac{di}{dt} &= \beta\kappa si - \gamma i \\
\frac{dr}{dt} &= \gamma i \\
r &= 1 - (s + i)
\end{aligned}$$

In these equations, $s$, $i$, and $r$ are now the proportions of the population in those categories: $s = \frac{S}{N}$, $i = \frac{I}{N}$, and $r = \frac{R}{N}$. Using the equations above and assuming all contacts are equally likely, the change in number of infectious individuals is related to the size of the susceptible group, the size of the infectious group, number of contacts, and parameters of the infection. Individuals leave the susceptible group at the rate they become infected and individuals

leave the infectious group at the rate they recover. These equations are deterministic in that the progress of the epidemic is completely determined by the disease parameters. Compartment models can be used to predict the maximum and cumulative incidence and to test mitigation strategies.[105] Also, given the number of infected individuals in an actual epidemic, it is possible to estimate the probability of transmission using these models.

Traditionally, mean-field models assume random mixing of the population, whereby each person is equally likely to form a contact with another in the population at each time point. Contacts are considered at the population level without regard to which individuals actually have contact with each other. If all contacts are random, the proportion of susceptibles that become infected is directly related to the proportions of infected and susceptible individuals in the population. In real populations, individuals are not equally likely to form a contact with any other individual in the population. Contacts in real networks are decidedly non-random and with some individuals having many more contacts than others. Modifications include classifying subgroups which can have different probabilities of contacts within and between subgroups.[106] Given $\gamma$, these equations can be modified to produce an effective $\beta$ that includes non-uniform mixing.[107] However, even this correction does not adequately represent the connections between individuals in real communities. Another assumption is that the population, $N$, is very, very large.[4] This assumption becomes untenable when examining disease transmission in small groups like classrooms.

### 5.1.1 Disease Spread in Networks

In order to more completely examine the possible routes of infection spread in a community, the $SIR$ model of disease spread has been applied to networks, creating a field called *network epidemiology*. Networks that examine transmission of disease have been called *risk-potential* networks.[108]. Using networks it is possible to more particularly take into consideration the location and contacts of individuals within a community during an epidemic. In this framework, the $SIR$ process is local compared to the global process in mean-field equations. Unlike in population-based models where all susceptibles in a population risk infection from the infectious individuals, only those susceptibles with a direct tie to an infectious individ-

ual are at-risk for infection. This replaces the random-mixing assumption because each individual has defined contacts through which they can transmit or receive infection.[102] This also allows for an examination of the effect of network position in epidemic outcomes (see Centrality section below). Networks also allow for the examination of disease spread in finite populations, a more realistic assumption for many situations. A network model of disease transmission also allows the infection begin simultaneously in different parts of the network compared to a traditional model which does not incorporate location.[103] While population-based $SIR$ models can be used to identify the course of an outbreak or epidemic, networks can illustrate the underlying transmission structure of the outbreak.[3] The network approach to studying disease transmission is especially useful for infections with a low prevalence or for local outbreaks of disease.[109] Unsurprisingly, the predictions between mean-field models and networks models have been shown to be very different.[110, 111]

In the network $SIR$ model, the definition of and movement between disease compartments ($S$, $I$, $R$) are the same as in population-based models but the process of compartment change is slightly different. Now the infection rate, $\beta$ is the rate per unit time that the infection will be transmitted from one infected individual to a susceptible individual through their contact (edge). Thus, $\beta$ is now conditional on contact between a susceptible and an infectious individual. The rate per unit time for which the infected individual remains infected is $\gamma$.[1] $SIR$ epidemics that occur in a closed network eventually end with extinction of the infection because the number of susceptible nodes is finite.[112] The corresponding network for a mean-field model is a regular network where all nodes have the same number and structure of connections. Then, "the dynamics of infection depend only upon *how many* nodes are infected rather than *which* particular nodes are infected."[111]

**Real Disease Networks** Network analysis methods have been applied to further examine known disease outbreaks in real networks using $SIR$ and related compartment models.[113, 25, 114] One subject area where network research is popular is in the study of sexually transmitted infections. Sexual contact networks are often determined for public

health interventions and so easily lend themselves to network analysis. For example, network analysis of a gonorrhea outbreak in Canada suggested that important individuals in propagating the disease could be identified by information centrality.[60] A study investigating the sexual network of adolescents found the structure of the network associated with disease transmission.[61] A network study of HIV-positive individuals in Chicago, IL suggested that bridging between groups influenced disease transmission.[41] Another study of chlamydia and gonorrhea in Canada examined component size and structure over two time periods and found the distribution of component size to be similar but the actual size and structures to be different.[115]

Alternatively, network analysis methods have been used to simulate potential disease outbreaks in real networks. Complicated network models for disease simulation that incorporate household and community contacts across an entire city have been explored in the *EpiSims* and *EpiSemdemics* programs.[116, 117, 118] The EpiSims model was used to determine the best public health strategy (such as targeted/limited/mass vaccination, quarantine, closing of malls/schools/churches, etc.) for mitigating infectious disease spread in an urban area. EpiSims utilizes census, land-use, and transportation data for contacts and disease state transition is probabilistic for each individual in the model. The EpiSims simulations are conducted instead of a differential equations approach. EpiSemdemics is a similar model but can be implemented on a larger scale, like a state or region.[118] Another similar model utilizing census and transportation data for contacts evaluated vaccination strategies for the city of Toronto, Canada.[119] Similarly, the effect of the airline transportation network has been shown to be important for international disease transmission.[120]

**Centrality** Node centrality is important for disease spread. Studies have shown that the largest number of cumulatively infected nodes occurs when the most central node, according to a variety of different measures, is infected first. For example, this exact result was found in a study of the all-around centrality measure (a combination of betweenness, degree, and k-core centrality) in both the Enron email network and the high energy physics community. In the Enron network, infecting the node with the highest degree centrality produces

more cumulatively infected nodes than randomly infecting any node, although less than infecting the all-around node. Similarly, in the high energy physics network, infecting the node with the highest betweenness centrality produces more cumulatively infected nodes than randomly infecting any node, although less than infecting the node with the highest degree centrality (which is the all-around node in this network).[25] Unfortunately, there are no measures of uncertainty or significance with this analysis. The correlation between the cumulative number of infected nodes at 10 days and closeness, betweenness, degree, and local centrality was examined for blogs on MSN Spaces, router-level topology of the internet, and network science co-authorship. Interestingly, the relationship between cumulative number of infected nodes at 10 days and centrality depends on the type of network. Using a scatter plot, closeness and local centrality have a positive correlation for MSN blogs, netscience co-authorship, and email network but not for blogs. In contrast, degree and betweenness centrality are not correlated with the cumulative number of infected nodes.[33] In opinion networks, beginning the epidemic in nodes highly ranked by LeaderRank centrality measure produces an epidemic that cumulatively infected more nodes than if the epidemic began with nodes highly ranked by PageRank.[121] Random walk betweenness, betweenness, farness (the inverse of closeness), and degree centrality have also been shown to be associated with time to infection and risk of infection in small world and random networks.[122] Hubs, or nodes with very high degree relative to the rest of the network, are important for transmission and network resilience.[1]

Nodes with high core index values in the network, as defined by the k-shell decomposition method, produce an epidemic that cumulatively infected more nodes compared to nodes of the same degree but with lower core index values.[113, 123] Core index value is related to the density of connections around a node, suggesting that in addition to position within the network, the actual network structure could be important in disease transmission. Additional relevant literature includes identifying nodes that might be extremely important in spreading disease in the network.[114, 25, 113]. Other research has developed a network method to estimate the number of infected individuals in an epidemic based on an important, or *superspreader*, node.[114]

The practical importance of centrality in disease spread is clearly illustrated by a study that monitored the outbreak of H1N1 influenza among Harvard undergraduate students.[124] The authors randomly selected students to be in the study and then asked the randomly selected students to nominate a friend to be in the study as well. The authors found that the nominated students were more central in the Harvard undergraduate student network and also experienced flu earlier than the randomly selected students. The results of this study suggest that centrality is not only theoretically important for disease spread, but also important in real world situations where it could provide a method of early detection in networks. This study also highlights that understanding network properties can provide new insights into epidemic behavior in populations.

Although it has been stated that "the initial spread and long-term behavior of any infectious disease are determined by both its epidemiological characteristics and the graph theoretical properties of the network", no research has been conducted on disease spread and centralization[125] Recall that network epidemiology examines the importance of contacts between and the location of individuals in a network on an epidemic process. Centralization, which is a global measure of how much one node dominates a network, is defined by the structure of the network and so could be very important for disease spread. The goal of this study is to statistically evaluate disease spread in the context of network centralization. Additionally, the current study will improve upon previous research by utilizing statistical methods to determine the effect of initially infecting a randomly selected node or the most central node. The relationship between different centrality/centralization measures and important epidemiologic endpoints will be also be investigated.

**Simulations** Indeed, much research has been conducted simulating disease spread in networks of different types, including random networks, lattice or regular networks, small world networks, and preferential attachment networks.[126, 102, 122, 103, 127, 128, 129, 112, 90, 110, 111]. Mathematical models of disease spread have been evaluated in random graph models.[130, 111] $SIS$ models have been investigated in networks where the importance of contacts between nodes and directional contact is allowed to vary by creating directed

edge weights.[131] The effect of mixing in a regular network has been examined and it has been shown that the network must have a large amount of mixing to produce results similar to those obtained by differential mean-field equations.[107] Some work has examined $SIR$ models in regular networks with varying amounts of mixing.[107] Simulations of the SARS outbreak on a variety of network models, including small-world type networks, random networks, and networks with a truncated power law degree distribution were conducted.[64, 65]

Parasite models have also been examined in lattice networks with particular emphasis on the relationship between connectivity of the network and parasite virulence.[110] Network analysis has also been used to examine transmission of an economic crisis between countries that are economically linked through international companies and trade relations.[123] Additionally, network analysis methods suggest that computer viruses can continue to circulate at low levels indefinitely.[128, 129] The effect of the airline transportation network has been shown to be important for international disease transmission.[120] The $SIR$ model has even been modified to simulate rumor spreading in networks.[132]

Disease spread using $SIR$ models has also been simulated in real networks, such as the Enron email network, high energy physics citation network, network science co-authorship, blogs on MSN Spaces, router-level topology of the internet, various email networks, and the *delicious.com* website.[25, 33, 121, 113] When simulating a disease spread process on a real network, often many simulations are done varying which and how many node(s) is/are infected first. In the case of the Enron email system and high energy physics citation network, 1,000 simulations were conducted for each type of simulation (random or most central node infected first).[25] Often, 100 simulations were conducted for each network.[33, 133] In a study of the spread on an economic crisis 50 simulations were conducted.[123] In these cases, the average over the simulations is used.[25, 33, 123, 133] Often these simulations are to demonstrate the effectiveness of a new centrality measure.

Of course, critical parameters in stochastically simulating a $SIR$ process on a network are the probabilities of transmission and recovery. Unfortunately, this information is difficult to obtain in a network setting as it requires knowledge of the disease at an individual

level instead of the population level. In one small world and percolation theory study the probability of transmission investigated was quite high, ranging from 0.4 to 0.6 (since an $SIS$ model was used no recovery probability was used).[127] For $SIS$ simulations on a network of Oregon router views, $\beta = 0.14$ and $\gamma$ was 0.08 or 0.24.[112] Parameter values for $SIS$ simulations in Barabási-Albert Preferential Attachment models were $\beta$ equal to 0.125, 0.15, 0.175 and $\gamma = 0.8$.[112] Parameter values for $SIS$ simulations in Erdös-Rènyi random graphs were $\beta = 0.2$ and $\gamma$ equal to 0.24,0.48, 0.72.[112] Another $SIR$ study in Erdös-Rènyi random graphs, Barabási-Albert Preferential Attachment graphs, a high school interaction network, and the network of contacts in Portland, OR used $\beta = 0.1$ and $\gamma = 0.2$.[133] Another simulation study for $SIR$ and $SIS$ models of sexually-transmitted disease spread investigated the ratio of transmission probability to recovery probability between 0.0 and 0.6.[125] A simulation study for computer viruses in continuous time following an $SIS$ model used transmission rates of 0.065 and 0.1.[128] Another continuous time study used probability of transmission in an $SIR$ model of disease spread in small world and random networks of 0.00375, 0.0075, and 0.015.[122] In an $SIR$ simulation on the Enron email network $\beta = 0.01$ and $\gamma = 0.3$.[25] An $SIR$ simulation of infection the *delicious.com* network used 0.5 as the probability of transmission for one randomly chosen contact of an infected node and probability of recovery related to the degree of the network.[121] Other studies have deliberately used small probabilities of transmission ($\beta = 0.01 - 0.08$) to limit the size of the epidemic.[113]

**Mathematical Description of Network Transmission Theory** A mathematical framework for the progress of epidemics in populations can be described using ordinary differential equations or Markov chain theory. As described above, these methods require important assumptions such as a very large population, random mixing, and equal probability of contacts between individuals. Network analysis allows an examination of disease spread in populations where there are a finite number of people and the number and type of contacts varies for each individual. Due to the uniqueness of each individual in the network, ordinary differential equations cannot be applied to networks when there is no uniform structure across the network. Markov chain theory has been successfully applied to networks.[126,

130, 134, 105] Early uses of Markov chain theory in networks considered Erdös-Rènyi $Gnp$ random graphs to be a Reed-Frost chain-binomial process.[126, 130] Infection in this model corresponds to the probability of randomly adding an edge between an infected node and any of the susceptibles in the network. Using a random graph framework, the number of nodes in the largest component can be calculated which could be interpreted as the size of the epidemic. Unfortunately, this approach is limited to random graphs which limits examination of network properties like centralization (see [97]). An exact Markov chain model for any network of $n$ nodes would have $2^n$ ($SI$ model) or $3^n$ ($SIR$ model) states, making computation impossible for large graph orders. More recently, N-intertwined Markov chain models have been suggested that reduce the number of possible states of the chain.[134, 105] Although Markov chain theory can describe epidemics in networks, it is not a reasonable method to investigate the association between network properties and disease spread due to the intense computing requirements.

An alternative approach is to stochastically simulate epidemics on networks and derive results based on the findings. This approach has been commonly employed for research investigating the influence of node-level properties on disease spread (see [25, 121, 33, 113] and others).

### 5.1.2 Epidemic Threshold

In infectious disease epidemiology, the *basic reproductive number* or *basic reproduction ratio*, $R_0$, is traditionally defined as the average number of secondary infections produced when one infected individual is introduced into a host population where everyone is susceptible.[135] However, $R_0$ has many different definitions with subtle distinctions between them. The basic reproductive number has also been defined as:

1. The expected number of new cases generated by a typical infectious individual in a large, susceptible population[136]
2. The number of people in a susceptible population that are directly infected by the introduction of a single infective[116]

3. The average number of individuals directly infected by an infectious case during his or her entire infectious period, when he or she enters a totally susceptible population[104]

4. The expected number of new infectious hosts that one infectious host will produce during his or her infectious period in a large population that is completely susceptible[137]

5. The average number of secondary infected individuals when a single individual is infected individually.[105]

All of these definitions assume a large, homogeneously mixing susceptible population. In an epidemic setting, $R_0$ is calculated by contact tracing of infectious contacts. All contacts of the initially infected individuals are followed and then tested to determine if the contacts secondarily contracted the infection. $R_0$ is then directly calculated by averaging the number of new infected cases produced by all of the initially infected individuals whose contacts were traced.[138] Using mean-field equations, where $\beta$ and $\gamma$ are known, can determine that $R_0 = \frac{\beta}{\gamma}$.

It has been noted that "$R_0$ is a convolution of transmission rate and contact patterns".[116] More specifically, the basic reproductive number depends on the probability of transmission from one infectious individual to a susceptible individual, the frequency of contacts, the duration of infectivity in an infectious individual, and the proportion of immune individuals in the population.[104] As a result, "the concept of $R_0$ finds its greatest use in the description of diseases that are spread broadly among individuals meeting more or less at random."[104]

Commonly $R_0$ is used as a *threshold criterion* where a $R_0 < 1$ produces a small epidemic that will eventually die out and $R_0 > 1$ produces a much bigger epidemic where a large portion of the population becomes infected.[136, 104, 137] When $R_0 < 1$, on average each infected individual infects less than one individual causing the epidemic to die out. On the other hand, when $R_0 > 1$ on average each infected individual infects more than one individual and the epidemic increases in size. Thus, $R_0$ can be considered an epidemic threshold and the terms are often used interchangeably. Using the traditional definition of

$R_0$ in a network setting, $R_0 = \langle k \rangle \frac{\beta}{\gamma}$, where $\langle k \rangle$ is the average degree of the network.[105] Of course, many different network structures with very different degree distributions can have the same average degree so the direct application of $R_0$ to networks is of limited utility.

As a result, in networks $R_0$ is replaced by the concept of epidemic threshold, $\tau$, which is compared to the *spectral radius* of the network. The spectral radius of a network is the first eigenvalue, $\lambda$, of the network's adjacency matrix. In a non-network setting, the spectral radius can be computed on the *disease transmission matrix.*[139]. Recall that spectral radius can only be computed for a connected network. Nold suggested that similar to $R_0$, if the spectral radius $> 1$ then the infection can persistent in the population.[139] However, more recent results suggest that for $SIS$ models $\tau = \frac{\gamma}{\beta}$ and $\lambda < \frac{\beta}{\gamma}$ produce an epidemic and $\lambda > \frac{\beta}{\gamma}$ produces an infection that dies out at least exponentially over time regardless of the number of nodes initially infected.[140, 112] This threshold has been confirmed by other researchers.[134, 105, 141] It has been shown that in preferential attachment the spectral radii of the networks are so small that infections can persist even at low levels of transmission.[128, 129] The expected number of infected nodes in an $SIR$ model is $\beta \frac{\lambda}{\gamma}$.[112] Thus, the size of an epidemic is a property of both the disease parameters ($\beta$ and $\gamma$) and the spectral radius of the graph. Importantly, the epidemic threshold $\tau$ is directly based on network structure, so is applicable to any undirected graph. Rigorous proofs for star graphs with $n$ nodes confirm that an $SIS$ epidemic dies out when $\tau = \sqrt{n} \leq \frac{\beta}{\gamma}$ and continues otherwise.[142]

## 5.2 Methods

### 5.2.1 Graph Generation

A modified version of the Star Start program is used to generate the networks for the SIR simulation.[97] The original Star Start method starts with a star graph, $g$, with $n$ nodes and then edges were randomly added or deleted (with equal probability) from the graph until $g$

was complete or empty. Loops and duplicate edges were not allowed. The modified version of Star Start restricts the number of graphs produced for each iteration to a maximum possible number of 3*$n$. Increasing the probability of adding an edge to 0.54 makes it slightly more likely that graphs with higher centralization levels are produced. These changes attempt to overcome two of the most important limitations of the Star Start program: length of time to iteration completion and the limited numbers of graphs produced with high centralization levels. Recall that the Star Start program does not have a restriction to only produce connected graphs. Consequently, a graph generated by the Star Start program may have more than one component, such as a disconnected node. A number of measures are collected for each network: density, number of components/clusters, number of nodes in the largest component, as well as closeness, betweenness, degree centrality and centralization. The largest component is identified for analysis and spectral value obtained. The Star Start graph generating method is repeated 500 times to ensure a range of graph structures and centralization values for analysis. Relative centrality and centralization measures are used to aid comparisons between different graph orders.[6] A brief description of the measures is below. Note that centralization and centrality are calculated on the full network.

**Closeness** Let $d(n_i, n_j)$ be the number of edges in the geodesic between node $n_i$ and node $n_j$. By convention, if node $n_i$ and node $n_j$ are not connected by any edges, then $d(n_i, n_j) = \infty$. Of course, the number of edges between node $n_i$ and itself is 0, so $d(n_i, n_i) = 0$. Then the relative closeness centrality (called *closeness centrality* from here forward) of node $n_i$ (in a connected graph) is defined as $C'_c(n_i) = \left[\frac{\sum_{j=1}^{n} d(n_i,n_j)}{n-1}\right]^{-1} = \frac{n-1}{\sum_{j=1}^{n} d(n_i,n_j)}$. By definition, closeness centrality can only be computed for a connected graph (where each node is connected by at least one edge). However, *igraph* package in R substitutes $n$ for $\infty$ in the case of disconnected nodes so all nodes have closeness centrality values.[10, 11] Consequently, for this paper closeness centrality is computed for all graphs, connected or disconnected.

Relative closeness centralization (called *closeness centralization* from here forward) is: $C_c = \frac{\sum_{i=1}^{n} [C'_c(n^*) - C'_c(n_i)]}{(n^2-3n+2)/(2n-3)}$, where $C'_c(n^*)$ is the maximum closeness centrality of the observed

graph.

**Betweenness** Any nodes on the geodesic connecting two nodes are said to be between them. Betweenness centrality measures how often a node is between other nodes. $C_B(n_k) = \sum_{i}^{n}\sum_{<j}^{n} b_{ij}(n_k)$ where $b_{ij}(n_k) = \frac{g_{ij}(n_k)}{g_{ij}}$ is the number of geodesics connecting $n_i$ and $n_j$ that contain $n_k$ and $g_{ij}$ is the total number of geodesics connecting $n_i$ and $n_j$. Since the center point in a star graph obtains the maximum value, relative betweenness centrality (*betweenness centrality*) is defined as: $C'_B(n_k) = \frac{2*C_B(n_k)}{n^2-3n+2}$.

Relative betweenness centralization (*betweenness centralization*) is calculated as: $C_B = \frac{\sum_{i=1}^{n}[C'_B(n^*)-C'_B(n_i)]}{n-1}$ where $C'_B(n^*)$ is the maximum betweenness centrality of the observed graph.

**Degree** The number of edges connected to node $i$ is the degree of the node, $k_i$. Relative degree centrality is simply the degree count for each node standardized by the maximum possible degree in a graph with $n$ nodes, $n-1$.[6] More formerly, relative degree centrality (*degree centrality*) is defined as: $C'_D(n_k) = \frac{\sum_{i=1}^{n} a(n_i,n_k)}{n-1}$ where

$$a(n_i, n_k) = \begin{cases} 1, & \text{iff edge between } n_i \text{ and } n_k \\ 0, & \text{otherwise} \end{cases}$$

The range for relative degree centrality is 0 (disconnected node) to 1 (maximally connected node). Degree centrality is computed the same way for connected and disconnected graphs.

Relative degree centralization (*degree centralization*) is calculated as:
$C_D = \frac{\sum_{i=1}^{n}[C_D(n^*)-C_D(n_i)]}{n^2-3n+2}$ where $C_D(n^*)$ is the maximum degree centrality of the observed graph. The range of possible values is 0 to 1.

### 5.2.2 SIR Simulation

In a network setting, population size is now the number of nodes. Birth and death rates, which would change the number of nodes in the network, will not be considered in order to keep the network static. As a result, the model will only examine a closed population of individuals. Since the centrality/centralization measures of interest are not defined for dynamic networks, contacts between nodes will be kept constant. Infection can be transmitted to the immediate neighbors, or those nodes connected by an edge, of an infected node. Networks will be undirected thus assuming that contacts, and therefore infection transmission, can occur in any direction. Contacts will be unweighted, so all contacts are equally important for disease transmission. No restriction on connectedness in the network will be assumed so in some networks the infection may not be able to spread to every node if there are disconnected nodes. The time period will be considered one day.

The method implemented follows a simple $SIR$ network model for disease spread, where the probability of transmission and recovery are stochastic and computed for each node at each time point.[103, 143] The $SIR$ model is implemented in the full network as follows. All nodes are initially susceptible and one node is selected (either randomly or because it is the most central in the network) to become infected. Most central nodes are determined by finding the node(s) with largest value of the measure of interest (either closeness, betweenness, or degree). In the case of ties, one node is randomly selected from the group of nodes with the same maximum centrality value. From the initially infected node, the neighboring nodes at risk of infection are determined. Nodes are infectious at the next time period after being infected. At the next time period, a random number $r_i$ is chosen from the uniform distribution for each at risk node $n_i$ and used to represent the probability of transmission of infection from the infected node to each at risk susceptible node. If $r_i \leq \beta$, the probability of transmission, then $n_i$ becomes infected. Transmission from multiple infected contacts is assumed to be independent so that susceptible node $n_i$ with links to $k_i$ infected nodes is at risk of infection from each infected node for a total probability of infection of $1-(1-\beta)^{k_i}$. In the subsequent time periods after the time of initial infection, a random number $s_i$ is chosen from the uniform distribution for each infected node. If $s_i \leq \gamma$, the probability

of recovery, then node $n_i$ recovers from infection and becomes immune. Otherwise, $n_i$ remains infected and able to transmit the infection to neighboring at risk nodes. Thus, node transition from susceptible to infected and from infected to recovered is probabilistic. The simulation continues until there are no new infected nodes and all infected nodes have recovered. All nodes have the potential to move from susceptible to infected to recovered and once recovered they are immune from infection. For each network one simulation is conducted where a random node is infected first. Then, one simulation where the most central node is infected first is conducted for each type of centrality examined: closeness, betweenness, and degree. In the interest of computational efficiency, if a node is a maximum on more than one centrality measure, the simulation results are copied.

Disease spread in connected networks is considered by examining the largest component of each full network. In this case, the most central node simulations begin from the most central node in the largest component. In the random node case, a random node is selected and if it is in the largest component, the results are the same as for the full network. If the random node selected is not in the largest component, no results are reported.

The $SIR$ process described above is implemented 100 times on each network generated by the modified Star Start program. Thus, when the most central node for a particular measure is selected, 100 $SIR$ simulations are conducted starting from that most central node. For the random node first simulations, a random node is selected and then the $SIR$ simulation conducted. Thus, unlike in the most central node simulations, in the random node simulations the infection is not required to start at the same node. Note that since each simulation ends when no new nodes become infected, the length of time for each simulation is variable.

Total number of nodes infected (current and newly infected) and cumulative number of nodes infected are recorded at each time point of the simulation for both the full network and the largest component. The total number of time periods from the start of the infection until completion is also recorded. These outcomes are recorded for the random node simulations and for the central node simulations for each type of centrality.

### 5.2.3 Analysis

Although the number of currently infected nodes and the number of cumulatively infected nodes is collected for each day of a simulation, this study focuses on five outcomes that are epidemiologically important: epidemic duration, number infected at the peak of the epidemic, day of the peak (as measured from the start of the simulation), final cumulative number of infected nodes, and the day that the final cumulative number of nodes infected is reached (again, as measured from the start of the simulation). The analysis will focus on these endpoints instead of considering the incidence and cumulative incidence over the entire course of the epidemic. Additionally, this analysis will only consider the largest component of the graphs generated by Star Start so the epidemic threshold can be included in the analysis. Spectral radius of the largest component is calculated and compared to epidemic threshold to determine if it is above or below. In addition, the relationship between closeness centralization and the endpoints is investigated using only connected graphs, thereby removing the effect of imputation from the analysis. Note that for connected graphs the largest component of the graph is the same as the full network.

Since 100 $SIR$ simulations were conducted on each graph generated and thousands of graphs were generated by the modified Star Start program, the data were summarized by graph for analysis. For each graph $g$ and outcome of interest $x$, the average of the outcome was taken across all 100 simulations: $\overline{g(x)} = \frac{1}{100}\sum_{i=1}^{100} g(x_i)$. The new unit of analysis is summarized graph information $g(\bar{x})$, where $x$ is epidemic duration, peak number infected, day of peak, final cumulative number infected, or day when no new nodes were infected.

To aid description of the effect of centralization, centralization was categorized by quartile into low, low-moderate, moderate-high, and high categories. Graphs were categorized by centralization quartile, largest component order, and whether they were above or below the epidemic threshold. The average of the summarized graph information was calculated for each category and outcome $x$: $\bar{x_c} = \frac{1}{n_c}\sum_{i=1}^{n_c} \overline{g_c(x_i)}$, where $n_c$ is the number of graphs in each category and $\overline{g_{c,i}(x)}$ is the summarized graph information for graph $i$ in category $c$. In addition, the standard deviation was calculated for each category. Results were analyzed

separately by type of node infected first (random or most central).

To further provide insight into the association of centralization with the epidemic process, the daily mean total or cumulative number of infected nodes and 95% confidence interval were plotted by centralization quartile and epidemic threshold for each graph order. Calculation of the daily mean and confidence interval were conducted as described above. Results were analyzed separately by type of node infected first (random or most central). Plots by centralization quartile compare the effect of type of node infected first.

To determine the relationship between the peak and final cumulative number of nodes infected and centralization for all the graph orders examined, two linear regression models were constructed. Each outcome was the dependent variable and the independent variables were degree centralization (continuous), most central node infected first/random node infected first (reference: random node infected first), order of the largest component, and a centralization and most central node infected first interaction term. Because the full network is not being used for this analysis, the order of the largest component was included in the model to control for the varying number of nodes in each graph. The centralization interaction term was added to incorporate any additional effect of centralization when the most central node is infected first. Graphs below the epidemic threshold were excluded from this analysis due to their small numbers. As with the descriptive analyses, the summarized graph information was used for this analysis. Model fit was assessed through diagnostic plots and estimated model parameters and R-squared reported. A similar analysis was conducted for the epidemiologic endpoints involving time (epidemic duration, peak day, day of last new infection) using a Cox proportional hazards model.

All analyses were conducted using R version 3.0.1.[73]. Graph generation and centrality/centralization measures were computed using the *igraph* package.[72]

## 5.3 Results

Graphs of order 5 to 40 nodes were generated for analysis. Analysis was conducted on each graph order separately. For all $SIR$ simulations, $\beta = 0.3$ and $\gamma = 0.2$ with a corresponding epidemic threshold of $\frac{0.2}{0.3} = 0.667$. Since the network is unweighted, a $\kappa$=1 is assumed. Degree centralization is the primary measure of interest because number of contacts is important for the spread of many diseases. However, the results for closeness and betweenness are also reported. The largest component of the network is used for analysis in order to incorporate the idea of epidemic threshold.

### 5.3.1 Degree Centralization

The number of graphs in each degree centralization quartile as well as the total number of graphs generated is shown in Figure 5.1. The figure also describes the number of graphs above and below the epidemic threshold. The modified Star Start program produces very few graphs below the epidemic threshold, especially as graph order increases. As graph order increases, almost all of the graphs below the epidemic threshold are in the low centralization quartile. No graphs below the epidemic threshold are produced when the graph order is 40. Movie 5.1 illustrates the actual distribution of degree centralization values by graph order, regardless of epidemic threshold. The majority of graphs generated are moderately centralized, with relatively fewer lowly centralized graphs as graph order increases.

Movie 5.2 plots the daily average total number of infected nodes in the largest component by degree centralization quartile and epidemic threshold for a random node infected first. Movie 5.3 plots the daily average total number of infected nodes in the largest component by degree centralization quartile and epidemic threshold for the most central node infected first. Movie 5.4 plots the daily average total number of infected nodes in the largest component by type of node infected first and epidemic threshold within a degree centralization quartile. Similarly, Movies 5.5-5.7 plot the daily average cumulative number of infected nodes in the largest component by degree centralization quartile and epidemic threshold for a random node infected first, for the most central node infected first, and then by type of node

infected first and epidemic threshold within a degree centralization quartile. In all the movies, graphs below the epidemic threshold have much shorter epidemics with many fewer nodes infected and there is no effect of centralization quartile. As shown in Movies 5.2 and 5.3, there are very clear differences in the incidence curve by centralization quartile, regardless of which node is infected first. The incidence curve has a higher peak and reaches the peak earlier as centralization quartile increases. The separation between the incidence curves increases as graph order increases when the most central node is infected first while it remains constant when a random node is infected first. However, for graphs with 5 nodes the effect of centralization is small.

**Epidemic Duration** Table 5.1 describes the average epidemic duration for graphs that are above the epidemic threshold. As shown in Table 5.1, the duration of the epidemic lengthens as graph order increases and the epidemic is also longer when the most central node is infected first compared to a random node infected first. Epidemic duration is the longest for graphs with low-moderate centralization and the shortest for graphs with high centralization levels. There is no difference in epidemic duration by type of node infected first for graphs below the epidemic threshold. In both cases, the epidemic duration is around 6 days (5.99 days (standard deviation (SD) 1.07) for random node and 5.97 days (SD 0.45) for most central node infected first. The epidemic duration is much shorter for graphs below the epidemic threshold. As shown in Table 5.2, all of the variables considered in the linear regression model are significantly associated ($p<0.0001$) with epidemic duration. As centralization increases, epidemic duration decreases although this effect is mitigated when the most central node is infected first. The adjusted R-squared of the model is 0.7254.

**Peak Number Infected** Table 5.3 describes the peak number of infected nodes for graphs that are above the epidemic threshold. As shown in Table 5.3, the peak number of infected nodes increases as centralization quartile increases, regardless of which node is infected first. For any graph order and centralization quartile, more nodes are infected at the peak when the most central node is infected first. As graph order increases, the peak number of infected nodes also increases. There is no difference in peak number infected by type

of node infected first for graphs below the epidemic threshold. In both cases, the peak number infected is 1 (SD 0), which is just the initially infected node. The peak number infected is much smaller for graphs below the epidemic threshold. As shown in Table 5.4, all of the variables in the linear regression model are significantly associated ($p<0.0001$) with number of nodes infected at the peak of the epidemic. After adjusting for order of the largest component, as centralization increases the peak number of infected nodes increases and this increase is even greater if the most central node is infected first. The adjusted R-squared of the model is 0.8893.

**Peak Day** Table 5.5 describes the average day of the peak number of infected nodes for graphs that are above the epidemic threshold. As shown in Table 5.5, the day of the peak occurs earlier as centralization quartile increases regardless of which node is infected first. When the most central node is infected first, the day of the peak occurs earlier than if a random node is infected first. As graph order increases, the day of the peak occurs later. There is no difference in the day of the peak number infected by type of node infected first for graphs below the epidemic threshold. In both cases, the day the peak number infected nodes occurs is day 2 of the epidemic (SD 0), which is the day the first node becomes infected. The day of the peak number infected occurs much earlier for graphs below the epidemic threshold. As shown in Table 5.6, all of the variables in the linear regression model are significantly associated ($p<0.0001$) with the day of the peak number of infected nodes. As centralization increases, the day of the peak decreases and this decrease is even greater if the most central node is infected first. The adjusted R-squared of the model is 0.6570.

**Cumulative Number Infected** Table 5.7 describes the average final cumulative number of infected nodes for graphs that are above the epidemic threshold, respectively. As shown in Table 5.7, regardless of which node is infected first there is a slight trend where the final cumulative number of infected nodes increases as centralization increases from low to moderate-high and then decreases when centralization is at high levels. Although it should be noted that graphs with centralization values above the bottom quartile have about the

same number of cumulatively infected nodes while graphs in the lowest quartile have many fewer cumulatively infected nodes. At all centralization levels and graph orders, infecting the most central node first produces a larger cumulative number of infected nodes than when a random node is infected first. Final cumulative number of infected nodes increases as largest component order increases. There is no difference in final cumulative number infected by type of node infected first for graphs below the epidemic threshold. In both cases, the final cumulative number infected nodes is 1 (SD 0) which is just the initially infected node. The final cumulative number of infected nodes for graphs below the epidemic threshold is much smaller than for graphs above the epidemic threshold. As shown in Table 5.8, all of the variables in the linear regression model are significantly associated ($p<0.0001$) with final cumulative number of infected nodes. The adjusted R-squared of the model is 0.8790. The cumulative number of infected nodes is much larger when the most central node is infected first.

**Day Final Number Infected** Table 5.9 describes the day that the final cumulative number of infected nodes was reached for graphs that are above the epidemic threshold. As shown in Table 5.9, number of days until the last new node is infected decreases substantially as centralization increases when the most central node is infected first. A similar pattern is true when a random node is infected first, although the effect is not as large between quartiles. The length of time until the last new node is infected increases as graph order increases. There is no difference in day final cumulative number of infected nodes is reached by type of node infected first for graphs below the epidemic threshold. In both cases, the day the peak number infected nodes occurs is day 2 of the epidemic (SD 0), which is the day the first node becomes infected. The day of the final cumulative number infected occurs much earlier for graphs below the epidemic threshold. As shown in Table 5.10, all of the variables in the Cox proportional hazard model are significantly associated ($p<0.0001$) with day that the final number of cumulative infected nodes was reached. The risk of reaching the day of the last new node infected earlier is much greater as centralization increases.

### 5.3.2 Closeness

In order to carefully evaluate the relationship between closeness centralization and the epidemiologic endpoints, the following results are for graphs that are connected. The number of graphs in each closeness centralization quartile as well as the total number of graphs generated is shown in Figure 5.2. All connected graphs were above the epidemic threshold. The number of connected graphs in the low centralization quartile is initially low and continues to decrease as graph order increases. Only one graph with 40 nodes was connected and also in the low closeness centralization quartile. A similar decrease, although not as severe, is seen for low-moderate and moderate-high quartiles. Movie 5.8 plots the distribution of closeness centralization values by connected graph order. As shown in the movie, very few graphs have low or low-moderate centralization and the vast majority of the graphs are very highly centralized.

Movie 5.9 plots the daily average total number of infected nodes for each connected graph order and closeness centralization quartile and epidemic threshold for a random node infected first. Movie 5.10 plots the daily average total number of infected nodes for each connected graph order by closeness centralization quartile and epidemic threshold for the most central node infected first. Movie 5.11 plots the daily average total number of infected nodes for each connected graph order by type of node infected first and epidemic threshold within a closeness centralization quartile. Similarly, Movies 5.12-5.14 plot the daily average cumulative number of infected nodes for each connected graph order by closeness centralization quartile and epidemic threshold for a random node infected first, for the most central node infected first, and then by type of node infected first and epidemic threshold within a closeness centralization quartile. As shown in Movies 5.9 and 5.10, the incidence curves overlap for the top three quartiles and the curve for the low quartile is much smaller than the others, regardless of which node is infected first. As shown in Movie 5.11, the incidence curve has a higher peak and reaches the peak earlier when the most central node is infected first. The same pattern is seen in the cumulative incidence curves.

**Epidemic Duration** Table 5.11 describes the average epidemic duration for graphs that are above the epidemic threshold. When the most central node is infected first, the duration of the epidemic shortens as centralization increases. On the other hand, when a random node is infected first epidemic duration is about the same for the bottom three quartiles and then decreases when the centralization is high. Epidemic duration increases as graph order increases and is also longer when the most central node is infected first. As shown in Table 5.12, all of the variables considered in the Cox proportional hazard model are significantly associated ($p<0.0001$) with epidemic duration. As centralization increases, epidemic duration decreases although this effect is mitigated when the most central node is infected first.

**Peak Number Infected** Table 5.13 describes the peak number of infected nodes for graphs that are above the epidemic threshold. As shown in Table 5.13, when the most central node is infected first the peak number of infected nodes increases as closeness centralization quartile increases. When a random node is infected first, the peak number of infected nodes does increase as centralization increases from low to moderate-high and then decreases when centralization is high. As graph order increases, the peak number of infected nodes also increases. As shown in Table 5.14, all of the variables in the linear regression model are significantly associated ($p<0.0001$) with number of nodes infected at the peak of the epidemic. If the most central node is infected first, the peak number of infected nodes increases as centralization increases. If a random node is infected first, the peak number of infected nodes actually decreases as centralization increases. The adjusted R-squared of the model is 0.9272.

**Peak Day** Table 5.15 describes the average day of the peak number of infected nodes for graphs that are above the epidemic threshold. As shown in Table 5.15, the day of the peak occurs slightly earlier when the most central node is infected first, regardless of centralization quartile and largest component order. For both random node and most central node, the day of the peak occurs earlier as centralization increases. As connected graph order increases, the day of the peak occurs later. Parameter estimates and hazard

ratios produced by a Cox proportional hazard model of peak day are shown in Table 5.16. As shown in the table, increasing centralization increases the risk of earlier time to peak number of infected nodes especially when then most central node is infected first.

**Cumulative Number Infected** Table 5.17 describes the average final cumulative number of infected nodes for graphs that are above the epidemic threshold. As shown in Table 5.17, when the most central node is infected first the final cumulative number of infected nodes increases as centralization quartile increases from low to moderate-high although there is only a very small difference between low-moderate and moderate-high quartiles. Interestingly, graphs in the highest closeness centralization quartile have about the same number of nodes infected as graphs in the low quartile. A similar pattern is produced when a random node is infected first. At all centralization levels and graph orders, infecting the most central node first produces a larger cumulative number of infected nodes than when a random node is infected first. Final cumulative number of infected nodes increases as connected graph order increases regardless of centralization level or which node was infected first. As shown in Table 5.18, all of the variables in the linear regression model are significantly associated ($p<0.0001$) with final cumulative number of infected nodes. The adjusted R-squared of the model is 0.9247. The final cumulative number of nodes infected sharply decreases as centralization increases when a random node is infected first.

**Day Final Number Infected** Table 5.19 describes the day that the final cumulative number of infected nodes was reached for graphs that are above the epidemic threshold. As shown in Table 5.19, the day of the last new infection occurs earlier as centralization quartile increases, regardless of which node is infected first. There is a very large difference in the day that the final new node is infected between high closeness centralization quartile and all of the other quartiles. The day when no new nodes are infected increases slightly as graph order increases. There is very little difference between random and most central node infected first. As shown in Table 5.20, all of the variables in the Cox proportional hazard model are significantly associated ($p<0.0001$) with day of last new infection. The risk of reaching the day of the last new infection earlier increases substantially as centralization

increases, especially when the most central node is infected first.

### 5.3.3 Betweenness Centralization

The number of graphs in each betweenness centralization quartile and the total number of graphs generated are shown in Figure 5.3. The figure also describes the number of graphs above and below the epidemic threshold. Again, the modified Star Start program produces very few graphs below the epidemic threshold, especially as graph order increases. As graph order increases, almost all of the graphs below the epidemic threshold are in the low centralization quartile. No graphs below the epidemic threshold are produced when the graph order is 40. Movie 5.15 illustrates the actual distribution of betweenness centralization values by graph order, regardless of epidemic threshold. The majority of graphs generated are in the low-moderate and moderate-high centralization quartiles, with relatively fewer low centralization quartile graphs as graph order increases.

Movie 5.16 plots the daily average total number of infected nodes in the largest component by betweenness centralization quartile and epidemic threshold for a random node infected first. Movie 5.17 plots the daily average total number of infected nodes in the largest component by betweenness centralization quartile and epidemic threshold for the most central node infected first. Movie 5.18 plots the daily average total number of infected nodes in the largest component by type of node infected first and epidemic threshold within a betweenness centralization quartile. Similarly, movies 5.19-5.21 plot the daily average cumulative number of infected nodes in the largest component by betweenness centralization quartile and epidemic threshold for a random node infected first, for the most central node infected first, and then by type of node infected first and epidemic threshold within a betweenness centralization quartile. In all the movies, graphs below the epidemic threshold have much shorter epidemics with many fewer nodes infected and there is no effect of centralization quartile. As shown in Movies 5.16 and 5.17, there are very clear differences in the incidence curve between centralization quartiles, regardless of which node is infected first. The incidence curve has a higher peak and reaches the peak earlier as centralization quartile increases. The separation between the incidence curves increases as graph order

increases when the most central node is infected first while it remains constant when a random node is infected first. However, for graphs with 5 nodes the effect of centralization is small.

**Epidemic Duration** Table 5.21 describes the average epidemic duration for graphs that are above the epidemic threshold. As shown in Table 5.21, the duration of the epidemic is similar regardless of centralization quartile for graph orders less than 30. For graphs with at least 30 nodes, epidemic duration shortens as betweenness centralization quartile increases. When the most central node is infected first, the epidemic duration is longer for any graph order. As with degree centralization, there is no difference in epidemic duration by type of node infected first. In both cases, the epidemic duration is around 6 days (5.99 days (SD 1.07) for random node and 5.97 days (SD 0.45) for most central node infected first. The epidemic duration is much shorter for graphs below the epidemic threshold. As shown in Table 5.22, all of the variables considered in the Cox proportional hazards model are significantly associated ($p<0.0001$) with epidemic duration. As centralization increases, epidemic duration decreases although this effect is mitigated when the most central node is infected first.

**Peak Number Infected** Table 5.23 describes the peak number of infected nodes for graphs that are above the epidemic threshold. As shown in Table 5.23, the peak number of infected nodes increases significantly as centralization quartile increases and more nodes are infected at the peak when the most central node is infected first. As graph order increases, the peak number of infected nodes also increases. There is no difference in peak number infected by type of node infected first for graphs below the epidemic threshold. In both cases, the peak number infected is 1 (SD 0), which is just the initially infected node. The peak number infected is much smaller for graphs below the epidemic threshold. As shown in Table 5.24, all of the variables in the linear regression model are significantly associated ($p<0.0001$) with number of nodes infected at the peak of the epidemic. Specifically, the peak number of infected nodes increases as centralization increases and is even greater when the most central node is infected first. The adjusted R-squared of the model is 0.8652.

**Peak Day** Table 5.25 describes the average day of the peak number of infected nodes for graphs that are above the epidemic threshold. As shown in Table 5.25, the day of the peak occurs earlier as centralization quartile increases regardless of which node is infected first. Although the day of the peak occurs even earlier if the most central node is infected first. As graph order increases, the day of the peak occurs later. There is no difference in day peak number infected by type of node infected first for graphs below the epidemic threshold. In both cases, the day the peak number infected nodes occurs is day 2 of the epidemic (SD 0), which is the day the first node becomes infected. The day of the peak number infected occurs much earlier for graphs below the epidemic threshold. As shown in Table 5.26, all of the variables in the Cox proportional hazards model are significantly associated ($p<0.0001$) with the day of the peak number of infected nodes. The day of the peak occurs earlier if the most central node is infected and also when centralization increases.

**Cumulative Number Infected** Table 5.27 describes the average final cumulative number of infected nodes for graphs that are above the epidemic threshold. As shown in Table 5.27, when a random node is infected first many fewer nodes are infected cumulatively compared to when the most central node is infected first. Graphs with centralization values above the bottom three quartiles have about the same number of cumulatively infected nodes which is greater than the number of infected nodes for graphs in the highest centralization quartile. Final cumulative number of infected nodes increases as graph order increases. There is no difference in final cumulative number infected by type of node infected first for graphs below the epidemic threshold. In both cases, the final cumulative number infected nodes is 1 (SD 0) which is just the initially infected node. The final cumulative number of infected nodes for graphs below the epidemic threshold is much smaller than for graphs above the epidemic threshold. As shown in Table 5.28, all of the variables in the linear regression model are significantly associated ($p<0.0001$) with final cumulative number of infected nodes. The cumulative number of infected nodes is much larger when the most central node is infected first. The adjusted R-squared for the model is 0.8821.

**Day Final Number Infected** Table 5.29 describes the day that the final cumulative number of infected nodes was reached for graphs that are above the epidemic threshold. As shown in Table 5.29, number of days until the last new node is infected decreases as centralization increases regardless of which node is infected first. The day of the last new infection occurs slightly later if the most central node is infected but otherwise there is little difference between the two groups. The day when no new nodes are infected increases slightly as graph order increases. For all graph orders and starting the infection at either the random node or the most central node, the day the last node became infected is much earlier in graphs below the epidemic threshold. There is no difference in day final cumulative number of infected nodes is reached by type of node infected first for graphs below the epidemic threshold. In both cases, the day the peak number infected nodes occurs is day 2 of the epidemic (SD 0), which is the day the first node becomes infected. The day of the final cumulative number infected occurs much earlier for graphs below the epidemic threshold. As shown in Table 5.30, all of the variables in the Cox proportional hazard model are significantly associated ($p<0.0001$) with day that the final number of cumulative infected nodes was reached. The risk of reaching the day of the last new infection earlier increases as centralization increases.

## 5.4 Discussion

The modifications to the Star Start program successfully produced a large number of graphs across the full centralization range for all three centralization measures for all graph orders analyzed. For all three measures the distribution of centralization values is not uniform, although most centralization bins contain at least a small number of graphs. The distributions are increasingly skewed as graph order increases suggesting that further modifications to the Star Start program may be necessary to produce a more uniform distribution of the full range of centralization values for large graph orders. The form of the modifications to the Star Start program for this analysis were empirically derived based on preliminary results with the goal of producing graphs with the full range of centralization values for all three

centralization measures simultaneously. The success of the current results suggest that the distributional limitations observed can be further overcome by continuing to slightly change the probability of adding an edge and also by allowing the program to run slightly longer before completing an iteration. In the current analysis, the fact that most of the graph orders investigated met the graph generating limit instead of reaching a complete or empty graph suggests that the graph limit should be increased. Adjusting the modifications to Star Start to produce graphs across the full range for one centralization measure (instead of all three) would also improve performance.

Unfortunately, the modified Star Start program does not produce graphs across a range of spectral radii. As described in Figures 5.1-3, the program does not produce very many graphs below the epidemic threshold for the disease parameters chosen for this study. It is clear from the figures that graphs below the epidemic threshold are almost exclusively in the low centralization quartile. All graphs below the epidemic threshold had a spectral radius of zero. Preliminary analyses suggest that the majority of these graphs are lone nodes. A graph with order $> 1$ and zero spectral radius occurs when two nodes that are not mutually connected are themselves connected to the same set of nodes.[144] The distribution of spectral radius values by graph order for graphs above the epidemic threshold is shown in Figure 5.4. The median spectral radius value increases as graph order increases suggesting that the Star Start program is very unlikely to produce graphs below the epidemic threshold for large graph orders. Also, it is known that the minimum and maximum possible spectral radii for a graph are related to the number of nodes of the graph.[145] In fact, the maximum possible value for the spectral radius of a graph is $n - 1$.[144]

In addition, the program does not generate many connected graphs, especially in the lower closeness centralization quartiles. In part this result is expected since connected graphs are a subset of all possible graphs and the Star Start program was designed to efficiently produce a sample of graph structures with the full range of centralization values. Graphs with more than one component have necessarily lower centralization levels so connected graphs should generally have higher centralization levels than disconnected graphs.

For graphs above the epidemic threshold, all of the outcomes investigated (epidemic dura-

tion, peak number infected, day of peak number of infected nodes, final cumulative number of infected nodes, and day final cumulative number of infected nodes is reached) are significantly associated with betweenness, closeness, and degree centralization after adjusting for number of nodes in the largest component. Due to the very large sample size in this study, statistical significance should be considered jointly with clinical significance and meaningful trends in the data. In this regard, peak number of infected nodes, day of the peak number of infected nodes, and cumulative number of infected nodes have a clear relationship with centralization. Peak number infected and final cumulative number infected increase as centralization increases for betweenness and degree centralization. Meanwhile, day of peak occurs earlier as betweenness and degree centralization increases. Furthermore, if the most central node is infected first the effect of centralization is even greater. Based on the study results, the effect of centralization starts at about 10 nodes and seems to increase as graph order increases. The effect of centralization on disease spread in graphs with 5 nodes is small although this is probably due to the very limited size of the graph. Unsurprisingly, another strong factor is order of the largest component, which effectively limits the maximum size of the epidemic.

There does not seem to be a relationship between epidemic duration and day final new node infected and degree or betweenness centralization. Instead, graph order is the most important predictor for these endpoints. This finding seems reasonable as these endpoints are more strongly related to the disease parameters ($\beta$, probability of transmission and $\gamma$, probability of recovery) than the network structure.

Although other research has shown that more nodes become cumulatively infected when the most central node is infected first[25, 113, 123], this study is the first to examine the issue with statistical rigor. This study confirms that when a most central node is infected first, more nodes are infected at the peak and at the end of the epidemic compared to when a random node is infected first. In addition, this effect also depends on the centralization level of the network.

The relationship of closeness centralization with the epidemiologic endpoints was inconsistent and very different compared to betweenness and degree centralization. For example,

final cumulative number of infected nodes does not have a clear relationship with centralization but day of last new infection does. In connected graphs there was a linear relationship between closeness centralization and number infected at peak when the most central node was infected first but no relationship when a random node was infected first. A possible explanation for these results is that closeness is not an appropriate measure for disease spread and is capturing some other process. Indeed, a study by Borgatti investigating the flow of information in a network suggests that closeness is not a good measure for an infectious process.[13] The findings could also be related to the fact that the modified Star Start program produced many more low and low-moderate centralized graphs for closeness. Future research should more thoroughly examine the relationship between these important epidemiologic endpoints and closeness centralization.

Further analysis of the relationship between closeness centralization on all graphs (connected and disconnected) and the epidemiologic endpoints is described in the Appendix. The effect of the imputation can be seen by comparing the lower centralization quartiles of the tables in the Appendix to the tables in the Results section. The effect of closeness centralization is less clear when all graphs are considered. An analysis of only connected graphs for betweenness and degree centralization did not appreciably change the results.

This study provides further support for the idea that behavior of epidemics on networks can be described by an epidemic threshold that is based on the spectral radius of the network. For any graph order, there were many fewer infected nodes at the peak, a much shorter epidemic duration, and an earlier day of the peak compared to graphs above the epidemic threshold. These results suggest that the epidemic is not sustained when the spectral radius of the graph is below the epidemic threshold, regardless of the centralization of the network. Interestingly, differences on the outcomes were found whether the most central or random node is infected first as graph order increased. Specifically, results were independent of graph order when the most central node was infected but not when a random node was infected first. However, these results should be further investigated due to the small numbers of graphs below the epidemic threshold generated by the Star Start program.

In order to incorporate the idea of epidemic threshold, only the largest component of each

network generated by the modified Star Start program was used for this analysis. Results are similar between the full network and largest component when the most central node is infected first (not shown). This is because the most central node in the network is located in the largest component of the network. On the other hand, when a random node is infected first the results vary by whether or not the randomly chosen node is in the largest component of the network. If the node is in a small component with few nodes, then the epidemic is necessarily limited. If the randomly chosen node is located in the largest component, then the size of the epidemic is related to the spectral radius and epidemic threshold. Due to the confounding of component size and epidemic threshold, full network analysis when a random node was infected first was not pursued.

**Other Disease Parameters** In order to determine the sensitivity of the simulation to the disease parameters, more extreme $\beta$ values were tested. As described in the Methods section, 500 iterations of the same modified Star Start program were used to generate 30 node networks. Similarly, 100 $SIR$ simulations were conducted on each graph with the probability of recovery unchanged ($\gamma = 0.2$) and the probability of transmission ($\beta$) set at 0.7, 0.8, or 0.9. Plots of the incidence curves and cumulative incidence curves by degree centralization quartile and probability of transmission can be seen in Movies 5.29-5.34 in the Appendix. As shown in the movies, increasing the probability of transmission narrows the size of the incidence curve and also shifts the curve earlier in the epidemic. Importantly, whether the $\beta$ value is set to be a more realistic 0.3 or a very unlikely 0.9, the same effect of centralization is observed whereby the peak is earlier and higher as centralization increases. The effect of infecting the most central node first is also the same. These results suggest that centralization of a network is important regardless of the disease parameters that are chosen.

**Other Analytic Approaches** The extremely large numbers of graphs generated (see Figures 5.1, 5.2, and 5.3) combined with 100 $SIR$ simulations per graph produced a lot of data that was challenging to analyze. In contrast, other research examining disease spread in networks using simulations perform multiple simulations on only one network.[117] Each

simulation is independent on the graph that it is run on so the simulations are conditionally independent by graph. For example, for a 5 node network 6,474 graphs were generated by the modified Star Start program. One hundred $SIR$ simulations were conducted on each of these 6,474 graphs. The outcome of each simulation was the total (or cumulative) number of infected nodes for each day of the simulation, thus the outcome is longitudinal. The optimal analytic solution would be a model that controlled for variability in the outcome within a graph and the variability between graphs with different centralization values for a longitudinally measured outcome.

Several attempts were made to incorporate the likely correlation between $SIR$ simulations on the same graph. First, for the 5 node networks a simple generalized estimating equations (GEE) model with total number of infected nodes predicted by closeness centralization quartile and time was implemented using a Poisson distribution and a log link with 6,474 clusters. Centralization quartile was treated as a categorical variable with low centralization as the referent group. GEE analysis was conducted using the *geepack* package in R.[146, 147, 148] An exchangeable correlation matrix was assumed. Unfortunately, this model took one week to produce results on the high performance computing cluster. Consequently, this analytic approach was abandoned as too time-consuming. However, it should be noted that in this model, closeness centralization quartile was strongly associated with total number of infected nodes ($p < 0.0001$). Results from this model are shown in Figure 5.6.

Next, a GEE analysis with a gamma variance distribution and inverse link was conducted to determine if closeness centralization quartile was significantly associated with mean number of infected nodes over time using the summarized graph data. An exchangeable correlation matrix was assumed and each summarized graph is considered a unique observation. Centralization quartile was treated as a categorical variable with low centralization as the referent group. Since the domain of the gamma distribution requires real numbers greater than zero, an average of zero infections (on the first day of the simulation) was very slightly increased to 0.001 infections for this analysis. Separate models were fit for random node infected first and most central node infected first. A comparison of the parameter estimates between the Poisson model clustered by graph and the gamma model with aggregated graph

information is shown in Figure 5.6. As seen in the table, magnitude and direction of the effect of centralization quartile is the same between the models. Unfortunately, this analysis was limited to graphs of order 5-15 nodes due to computing limitations. Results from this attempt are shown in Figure 5.7.

**Limitations** Some limitations are due to the use of Freeman relative centrality/centralization measures. For example, these measures are not defined for weighted or dynamic networks. Unfortunately, an unweighted network will only capture whether or not any connection between two nodes (people) exists, not the exact duration of contact (which would provide a precise measure of exposure of susceptible contacts to infectious contacts). As a result, the probability of transmission is uniform across all contacts. For some infectious diseases, this limitation could be important if the probability of transmitting the infection is directly related to amount of contact. A modification to the current model, assuming that greater contact means a correspondingly greater probability of infection, is to vary the probability of transmission between pairs nodes to reflect the combined effect of contact and transmission.[6]

Similarly, the networks used for the $SIR$ simulations are static and as such do not add or remove nodes or edges. Addition and deletion of nodes in the network could be used to more closely emulate real populations where there are births, deaths, immigration, and emigration. In the context of an $SIR$ epidemic, a static population is a reasonable assumption because it is unlikely that the size of the population would change substantially over the several week period of the epidemic. Addition and deletion of edges reflects changing contact patterns, which could be important if contacts change more quickly than the infection is transmitted. Some research on dynamic networks suggests that adding more edges increases the number of nodes infected at the peak and also moves the peak earlier.[133]

Another limitation due to the measures themselves is that closeness and betweenness utilize shortest paths in their computation and as such assume that the contagion only spreads through the most direct route in the network. As a result, the networks generated by the Star Start program may not be appropriate for some types of disease transmission.

Previous research has suggested that sexually transmitted infections are more suited to tree-like networks which are unlikely to be generated by the Star Start program.[61]

Another assumption is that disease can be transmitted between any two nodes with contact (that have an edge between them). This would not be true if behavioral modifications were in place (wearing a mask, vaccination, condom usage, etc.) that allowed contact between individuals but did not allow disease transmission. The current method does not allow for disconnected nodes because only the largest component of the network is used for the $SIR$ simulation. A disconnected node means that a person with no contacts can not become infected by anyone and they also cannot infect others in the network. This is a reasonable assumption since it is likely that at least some small portion of individuals in a community will have little to no contact with others during an epidemic. Unfortunately, the epidemic threshold cannot be computed on disconnected networks and so the analysis excluded disconnected nodes and small components.

Other limitations are due to the intense computational requirements of the simulations. The large number of simulations per graph produced extremely large files and required several weeks to compute as graph order increased. For example, simulations for a 40 node graph took three weeks with resulting file sizes of about 1 GB each. Unfortunately, computer memory restrictions did not allow for the saving of infection path information, which would have enabled direct $R_0$ calculations based on the secondary case ascertainment definition.

**Future Directions** Future directions include simulating $SIS$ and $SIRS$ processes in small centralized networks to determine if the results are similar to those observed for $SIR$. If only one graph order is considered, then the influence of weighted contacts on the results could be investigated. Of course, performing the same analysis for larger networks should be considered if the computational limitations can be overcome.

## 5.5 Conclusion

Centralization is significantly associated with epidemic severity for all of the graph orders considered. Specifically, as degree or betweenness centralization increases the peak number of infected nodes increases, time until the peak decreases, and the final cumulative number of infected nodes also increases. Closeness centralization does not have as strong of a relationship and should only be considered for connected networks.

## 5.6 Figures, Movies, Tables

### 5.6.1 Degree Centralization

| Degree Centralization | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Graph Order (# nodes) | Number of Graphs | | | | | | | | |
| | Below Threshold | | | | Above Threshold | | | | |
| | Low | Low-Mod | Mod-High | High | Low | Low-Mod | Mod-High | High | Total |
| 5 | 144 | 49 | 10 | 0 | 1,418 | 2,439 | 1,488 | 1,016 | 6,564 |
| 10 | 152 | 39 | 0 | 0 | 2,832 | 5,696 | 3,479 | 2,572 | 14,770 |
| 15 | 80 | 0 | 0 | 0 | 3,741 | 7,965 | 5,947 | 4,675 | 22,408 |
| 20 | 39 | 0 | 0 | 0 | 4,551 | 10,919 | 8,417 | 6,074 | 30,000 |
| 25 | 31 | 0 | 0 | 0 | 5,302 | 14,133 | 10,176 | 7,850 | 37,492 |
| 30 | 11 | 0 | 0 | 0 | 5,190 | 17,542 | 13,137 | 9,120 | 45,000 |
| 35 | 17 | 0 | 0 | 0 | 5,411 | 20,389 | 15,808 | 10,875 | 52,500 |
| 40 | 0 | 0 | 0 | 0 | 5,080 | 24,505 | 17,787 | 12,628 | 60,000 |

Figure 5.1: Number of graphs generated by degree centralization category

Movie 5.1. Distribution of degree centralization values produced by the modified Star Start program for graphs of order 5-40 nodes.

(Loading Video...)

Movie 5.2. Daily average total number of infected nodes in the largest component by degree centralization quartile with random node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.3. Daily average total number of infected nodes in the largest component by degree centralization quartile with most central node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.4. Daily average total number of infected nodes in the largest component by degree centralization quartile and type of node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.5. Daily average cumulative number of infected nodes in the largest component by degree centralization quartile with random node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.6. Daily average cumulative number of infected nodes in the largest component by degree centralization quartile with most central node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.7. Daily average cumulative number of infected nodes in the largest component by degree centralization quartile and type of node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Table 5.1: Average epidemic duration by degree centralization quartile and largest component order for graphs above the epidemic threshold

| Degree Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Epidemic Duration Graphs Above Epidemic Threshold | | | | | | |
| Order | Cent Quartile | n | Random First | | Most Central First | |
| | | | Mean | SD | Mean | SD |
| 5 | Low | 1605 | 11.39 | 1.60 | 11.78 | 1.16 |
| | Low-Mod | 1725 | 11.87 | 1.07 | 12.22 | 0.83 |
| | Mod-High | 753 | 11.36 | 0.98 | 11.88 | 0.77 |
| | High | 1016 | 10.90 | 0.80 | 11.58 | 0.67 |
| 10 | Low | 1386 | 14.06 | 1.87 | 15.10 | 1.43 |
| | Low-Mod | 2646 | 14.85 | 1.48 | 15.85 | 1.04 |
| | Mod-High | 1386 | 14.65 | 1.42 | 15.67 | 0.94 |
| | High | 1645 | 13.35 | 1.09 | 14.92 | 0.76 |
| 15 | Low | 1240 | 15.87 | 2.06 | 17.50 | 1.51 |
| | Low-Mod | 3354 | 16.43 | 1.78 | 18.07 | 1.16 |
| | Mod-High | 2117 | 16.04 | 1.56 | 17.68 | 1.02 |
| | High | 2041 | 14.74 | 1.24 | 16.83 | 0.80 |
| 20 | Low | 1346 | 17.07 | 2.07 | 19.32 | 1.61 |
| | Low-Mod | 3930 | 17.48 | 1.83 | 19.63 | 1.15 |
| | Mod-High | 2515 | 16.97 | 1.72 | 19.05 | 1.04 |
| | High | 1884 | 15.46 | 1.21 | 18.02 | 0.74 |
| 25 | Low | 1086 | 18.01 | 2.09 | 20.87 | 1.52 |
| | Low-Mod | 4092 | 18.37 | 1.91 | 20.89 | 1.18 |
| | Mod-High | 3066 | 17.56 | 1.61 | 20.05 | 0.99 |
| | High | 2294 | 16.19 | 1.35 | 19.05 | 0.79 |
| 30 | Low | 808 | 19.36 | 2.07 | 22.42 | 1.53 |
| | Low-Mod | 4272 | 19.33 | 1.90 | 22.11 | 1.14 |
| | Mod-High | 3629 | 18.24 | 1.67 | 20.99 | 1.02 |
| | High | 2527 | 16.73 | 1.29 | 19.88 | 0.75 |
| 35 | Low | 463 | 20.55 | 2.13 | 23.95 | 1.39 |
| | Low-Mod | 3302 | 20.63 | 1.74 | 23.46 | 1.01 |
| | Mod-High | 2747 | 19.15 | 1.64 | 21.97 | 0.97 |
| | High | 3008 | 17.24 | 1.34 | 20.61 | 0.76 |
| 40 | Low | 11 | 24.26 | 1.22 | 26.80 | 0.81 |
| | Low-Mod | 415 | 22.58 | 1.37 | 24.90 | 0.79 |
| | Mod-High | 195 | 21.59 | 1.66 | 23.83 | 0.93 |
| | High | 1390 | 17.26 | 1.21 | 20.96 | 0.72 |

Table 5.2: Parameter estimates from Cox proportional hazards model for epidemic duration for graphs above the epidemic threshold

| **Degree Centralization** | | | | |
|---|---|---|---|---|
| **Epidemic Duration** | | | | |
| **Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 2.7819 | 0.0100 | 16.1493 | $< 0.0001$ |
| Most Central Node First (ref: Random) | -1.0234 | 0.0075 | 0.3594 | $< 0.0001$ |
| Order of Largest Component | -0.1718 | 0.0002 | 0.8421 | $< 0.0001$ |
| Centralization*Most Central Node First | -0.4366 | 0.0131 | 0.6463 | $< 0.0001$ |

Table 5.3: Average peak number of infected nodes by degree centralization quartile and largest component order for graphs above the epidemic threshold

| Degree Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Peak Number Infected Graphs Above Epidemic Threshold | | | | | | |
| Order | Cent Quartile | n | Random First | | Most Central First | |
| | | | Mean | SD | Mean | SD |
| 5 | Low | 1605 | 2.74 | 0.57 | 2.91 | 0.48 |
| | Low-Mod | 1725 | 2.93 | 0.38 | 3.17 | 0.31 |
| | Mod-High | 753 | 2.71 | 0.31 | 3.01 | 0.28 |
| | High | 1016 | 2.60 | 0.16 | 3.00 | 0.14 |
| 10 | Low | 1386 | 3.83 | 0.93 | 4.41 | 0.83 |
| | Low-Mod | 2646 | 4.50 | 0.83 | 5.22 | 0.70 |
| | Mod-High | 1386 | 4.63 | 0.72 | 5.49 | 0.57 |
| | High | 1645 | 4.30 | 0.40 | 5.41 | 0.30 |
| 15 | Low | 1240 | 4.76 | 1.13 | 5.82 | 1.07 |
| | Low-Mod | 3354 | 5.68 | 1.04 | 7.00 | 0.87 |
| | Mod-High | 2117 | 6.11 | 0.84 | 7.62 | 0.66 |
| | High | 2041 | 5.95 | 0.54 | 7.76 | 0.39 |
| 20 | Low | 1346 | 5.55 | 1.29 | 7.09 | 1.30 |
| | Low-Mod | 3930 | 6.84 | 1.26 | 8.74 | 1.08 |
| | Mod-High | 2515 | 7.60 | 1.05 | 9.74 | 0.78 |
| | High | 1884 | 7.54 | 0.63 | 10.06 | 0.41 |
| 25 | Low | 1086 | 6.26 | 1.40 | 8.29 | 1.41 |
| | Low-Mod | 4092 | 7.92 | 1.53 | 10.38 | 1.33 |
| | Mod-High | 3066 | 8.97 | 1.08 | 11.77 | 0.82 |
| | High | 2294 | 9.22 | 0.86 | 12.44 | 0.55 |
| 30 | Low | 808 | 7.38 | 1.63 | 9.76 | 1.59 |
| | Low-Mod | 4272 | 9.16 | 1.59 | 12.12 | 1.36 |
| | Mod-High | 3629 | 10.38 | 1.21 | 13.78 | 0.93 |
| | High | 2527 | 10.77 | 0.90 | 14.68 | 0.56 |
| 35 | Low | 463 | 8.51 | 1.81 | 11.47 | 1.86 |
| | Low-Mod | 3302 | 10.75 | 1.78 | 14.11 | 1.52 |
| | Mod-High | 2747 | 11.90 | 1.36 | 15.76 | 1.03 |
| | High | 3008 | 12.33 | 1.04 | 16.93 | 0.62 |
| 40 | Low | 11 | 10.93 | 0.71 | 13.69 | 0.65 |
| | Low-Mod | 415 | 14.09 | 2.19 | 17.56 | 1.92 |
| | Mod-High | 195 | 15.01 | 2.02 | 18.88 | 1.51 |
| | High | 1390 | 13.85 | 1.01 | 19.28 | 0.56 |

Table 5.4: Parameter estimates from linear regression model for peak number of infected nodes for graphs above the epidemic threshold

| **Degree Centralization** | | | |
|---|---|---|---|
| **Peak Number of Infected Nodes Graphs Above Epidemic Threshold** | | | |
| **Parameter** | **Estimate** | **SE** | **p-value** |
| Intercept | -1.3579 | 0.0071 | $< 0.0001$ |
| Centralization | 2.6966 | 0.0105 | $< 0.0001$ |
| Most Central Node First (ref: Random) | 1.2323 | 0.0084 | $< 0.0001$ |
| Order of Largest Component | 0.3360 | 0.0002 | $< 0.0001$ |
| Centralization*Most Central Node First | 2.4915 | 0.0147 | $< 0.0001$ |

Table 5.5: Average day of peak number of infected nodes by degree centralization quartile and largest component order for graphs above the epidemic threshold

| Degree Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Day Peak Number Infected Graphs Above Epidemic Threshold | | | | | | |
| Order | Cent Quartile | n | Random First | | Most Central First | |
| | | | Mean | SD | Mean | SD |
| 5 | Low | 1605 | 4.51 | 0.48 | 4.38 | 0.33 |
| | Low-Mod | 1725 | 4.62 | 0.27 | 4.35 | 0.24 |
| | Mod-High | 753 | 4.58 | 0.25 | 4.17 | 0.19 |
| | High | 1016 | 4.45 | 0.26 | 3.94 | 0.19 |
| 10 | Low | 1386 | 5.74 | 0.61 | 5.42 | 0.52 |
| | Low-Mod | 2646 | 5.80 | 0.41 | 5.27 | 0.38 |
| | Mod-High | 1386 | 5.61 | 0.39 | 4.88 | 0.32 |
| | High | 1645 | 5.10 | 0.34 | 4.38 | 0.21 |
| 15 | Low | 1240 | 6.50 | 0.72 | 6.06 | 0.64 |
| | Low-Mod | 3354 | 6.40 | 0.61 | 5.68 | 0.53 |
| | Mod-High | 2117 | 5.96 | 0.54 | 5.08 | 0.42 |
| | High | 2041 | 5.36 | 0.41 | 4.55 | 0.23 |
| 20 | Low | 1346 | 7.02 | 0.76 | 6.54 | 0.68 |
| | Low-Mod | 3930 | 6.71 | 0.66 | 5.85 | 0.57 |
| | Mod-High | 2515 | 6.11 | 0.60 | 5.14 | 0.45 |
| | High | 1884 | 5.36 | 0.38 | 4.57 | 0.20 |
| 25 | Low | 1086 | 7.41 | 0.81 | 6.90 | 0.67 |
| | Low-Mod | 4092 | 6.96 | 0.71 | 6.01 | 0.58 |
| | Mod-High | 3066 | 6.13 | 0.58 | 5.12 | 0.41 |
| | High | 2294 | 5.42 | 0.40 | 4.60 | 0.20 |
| 30 | Low | 808 | 7.92 | 0.82 | 7.34 | 0.74 |
| | Low-Mod | 4272 | 7.23 | 0.74 | 6.20 | 0.61 |
| | Mod-High | 3629 | 6.23 | 0.60 | 5.17 | 0.41 |
| | High | 2527 | 5.45 | 0.37 | 4.62 | 0.16 |
| 35 | Low | 463 | 8.37 | 0.83 | 7.75 | 0.68 |
| | Low-Mod | 3302 | 7.62 | 0.66 | 6.52 | 0.54 |
| | Mod-High | 2747 | 6.44 | 0.58 | 5.31 | 0.39 |
| | High | 3008 | 5.48 | 0.37 | 4.64 | 0.16 |
| 40 | Low | 11 | 9.87 | 0.43 | 8.77 | 0.42 |
| | Low-Mod | 415 | 8.11 | 0.53 | 6.89 | 0.49 |
| | Mod-High | 195 | 7.24 | 0.53 | 5.96 | 0.38 |
| | High | 1390 | 5.38 | 0.32 | 4.61 | 0.14 |

Table 5.6: Parameter estimates from Cox proportional hazards model for day peak number infected for graphs above the epidemic threshold

| **Degree Centralization** | | | | |
|---|---|---|---|---|
| **Day Peak Number Infected** | | | | |
| **Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 4.9663 | 0.0109 | 143.4992 | $< 0.0001$ |
| Most Central Node First (ref: Random) | 0.4866 | 0.0077 | 1.6268 | $< 0.0001$ |
| Order of Largest Component | -0.0874 | 0.0002 | 0.9163 | $< 0.0001$ |
| Centralization*Most Central Node First | 2.6559 | 0.0140 | 14.2379 | $< 0.0001$ |

Table 5.7: Average final cumulative number infected nodes by degree centralization quartile and largest component order for graphs above the epidemic threshold

| Degree Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Final Cumulative Number Infected Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 1605 | 3.61 | 0.78 | 3.83 | 0.63 |
| | Low-Mod | 1725 | 3.89 | 0.50 | 4.14 | 0.39 |
| | Mod-High | 753 | 3.58 | 0.42 | 3.91 | 0.35 |
| | High | 1016 | 3.35 | 0.23 | 3.81 | 0.17 |
| 10 | Low | 1386 | 5.95 | 1.48 | 6.81 | 1.28 |
| | Low-Mod | 2646 | 6.87 | 1.23 | 7.83 | 0.97 |
| | Mod-High | 1386 | 6.87 | 1.10 | 7.95 | 0.81 |
| | High | 1645 | 6.02 | 0.65 | 7.44 | 0.45 |
| 15 | Low | 1240 | 8.16 | 2.02 | 9.84 | 1.81 |
| | Low-Mod | 3354 | 9.36 | 1.78 | 11.30 | 1.38 |
| | Mod-High | 2117 | 9.47 | 1.45 | 11.56 | 1.07 |
| | High | 2041 | 8.60 | 0.96 | 11.04 | 0.67 |
| 20 | Low | 1346 | 10.12 | 2.47 | 12.77 | 2.35 |
| | Low-Mod | 3930 | 11.71 | 2.21 | 14.70 | 1.73 |
| | Mod-High | 2515 | 12.03 | 1.90 | 15.15 | 1.37 |
| | High | 1884 | 10.92 | 1.11 | 14.42 | 0.75 |
| 25 | Low | 1086 | 11.97 | 2.81 | 15.73 | 2.65 |
| | Low-Mod | 4092 | 14.02 | 2.74 | 18.07 | 2.17 |
| | Mod-High | 3066 | 14.35 | 1.96 | 18.56 | 1.42 |
| | High | 2294 | 13.48 | 1.54 | 17.99 | 1.07 |
| 30 | Low | 808 | 14.73 | 3.28 | 19.36 | 3.03 |
| | Low-Mod | 4272 | 16.70 | 2.97 | 21.80 | 2.32 |
| | Mod-High | 3629 | 16.90 | 2.26 | 22.15 | 1.64 |
| | High | 2527 | 15.83 | 1.56 | 21.42 | 1.06 |
| 35 | Low | 463 | 17.63 | 3.73 | 23.48 | 3.48 |
| | Low-Mod | 3302 | 20.23 | 3.17 | 26.24 | 2.43 |
| | Mod-High | 2747 | 19.88 | 2.49 | 26.04 | 1.76 |
| | High | 3008 | 18.28 | 1.82 | 24.92 | 1.19 |
| 40 | Low | 11 | 24.88 | 1.40 | 30.81 | 0.95 |
| | Low-Mod | 415 | 26.73 | 3.33 | 32.97 | 2.33 |
| | Mod-High | 195 | 26.52 | 3.65 | 32.89 | 2.65 |
| | High | 1390 | 20.22 | 1.74 | 27.98 | 1.16 |

Table 5.8: Parameter estimates from linear regression model for final cumulative number of infected nodes for graphs above the epidemic threshold

| **Degree Centralization** | | | |
|---|---|---|---|
| **Final Cumulative Number of Infected Nodes Graphs Above Epidemic Threshold** | | | |
| **Parameter** | **Estimate** | **SE** | **p-value** |
| Intercept | -1.4454 | 0.0126 | $< 0.0001$ |
| Centralization | -0.6488 | 0.0187 | $< 0.0001$ |
| Most Central Node First (ref: Random) | 2.6141 | 0.0150 | $< 0.0001$ |
| Order of Largest Component | 0.6347 | 0.0004 | $< 0.0001$ |
| Centralization*Most Central Node First | 2.6100 | 0.0261 | $< 0.0001$ |

Table 5.9: Average day final cumulative number of nodes infected by degree centralization quartile and largest component order for graphs above the epidemic threshold

| Degree Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Day Final Cumulative Number Infected Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 1605 | 5.68 | 0.72 | 5.65 | 0.52 |
| | Low-Mod | 1725 | 5.83 | 0.38 | 5.62 | 0.32 |
| | Mod-High | 753 | 5.75 | 0.33 | 5.45 | 0.23 |
| | High | 1016 | 5.42 | 0.35 | 5.05 | 0.26 |
| 10 | Low | 1386 | 7.86 | 0.94 | 7.95 | 0.79 |
| | Low-Mod | 2646 | 7.86 | 0.66 | 7.76 | 0.59 |
| | Mod-High | 1386 | 7.48 | 0.65 | 7.18 | 0.55 |
| | High | 1645 | 6.46 | 0.59 | 6.15 | 0.50 |
| 15 | Low | 1240 | 9.28 | 1.17 | 9.60 | 0.96 |
| | Low-Mod | 3354 | 9.08 | 0.99 | 9.15 | 0.84 |
| | Mod-High | 2117 | 8.30 | 0.98 | 8.13 | 0.88 |
| | High | 2041 | 7.00 | 0.83 | 6.77 | 0.75 |
| 20 | Low | 1346 | 10.29 | 1.24 | 10.95 | 1.04 |
| | Low-Mod | 3930 | 9.86 | 1.09 | 10.12 | 0.94 |
| | Mod-High | 2515 | 8.81 | 1.12 | 8.76 | 0.99 |
| | High | 1884 | 7.05 | 0.85 | 6.88 | 0.81 |
| 25 | Low | 1086 | 11.09 | 1.36 | 12.05 | 1.05 |
| | Low-Mod | 4092 | 10.55 | 1.19 | 10.97 | 1.01 |
| | Mod-High | 3066 | 9.12 | 1.10 | 9.21 | 0.99 |
| | High | 2294 | 7.25 | 0.94 | 7.14 | 0.88 |
| 30 | Low | 808 | 12.07 | 1.35 | 13.18 | 1.09 |
| | Low-Mod | 4272 | 11.28 | 1.23 | 11.84 | 1.01 |
| | Mod-High | 3629 | 9.58 | 1.19 | 9.78 | 1.06 |
| | High | 2527 | 7.43 | 0.95 | 7.38 | 0.93 |
| 35 | Low | 463 | 13.02 | 1.38 | 14.36 | 1.06 |
| | Low-Mod | 3302 | 12.21 | 1.11 | 12.83 | 0.92 |
| | Mod-High | 2747 | 10.28 | 1.15 | 10.57 | 1.01 |
| | High | 3008 | 7.64 | 1.02 | 7.67 | 1.03 |
| 40 | Low | 11 | 16.07 | 0.94 | 16.68 | 0.75 |
| | Low-Mod | 415 | 13.20 | 0.85 | 13.50 | 0.97 |
| | Mod-High | 195 | 11.90 | 0.87 | 11.97 | 0.65 |
| | High | 1390 | 7.13 | 0.87 | 7.07 | 0.91 |

Table 5.10: Parameter estimates from Cox proportional hazards model for day peak number infected for graphs above the epidemic threshold

| **Degree Centralization** | | | | |
|---|---|---|---|---|
| **Day Final Cumulative Number Infected** | | | | |
| **Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 5.6106 | 0.0112 | 273.3060 | $< 0.0001$ |
| Most Central Node First (ref: Random) | -0.6181 | 0.0074 | 0.5390 | $< 0.0001$ |
| Order of Largest Component | -0.1532 | 0.0002 | 0.8579 | $< 0.0001$ |
| Centralization*Most Central Node First | 1.0966 | 0.0132 | 2.9939 | $< 0.0001$ |

### 5.6.2 Closeness Centralization

| Closeness Centralization | | | | | |
|---|---|---|---|---|---|
| Graph Order (# nodes) | Number of Connected Graphs Above Threshold | | | | |
| | Low | Low-Mod | Mod-High | High | Total |
| 5 | 628 | 1,190 | 963 | 1,469 | 4,250 |
| 10 | 232 | 2,031 | 1,435 | 1,859 | 5,557 |
| 15 | 142 | 1,791 | 1,534 | 2,163 | 5,630 |
| 20 | 97 | 1,299 | 1,176 | 1,788 | 4,360 |
| 25 | 33 | 724 | 792 | 1,758 | 3,307 |
| 30 | 42 | 621 | 656 | 1,464 | 2,783 |
| 35 | 1 | 455 | 547 | 1,477 | 2,480 |
| 40 | 4 | 282 | 332 | 1,393 | 2,011 |

Figure 5.2: Number of graphs generated by closeness centralization category

Movie 5.8. Distribution of closeness centralization values produced by the modified Star Start program for connected graphs of order 5-40 nodes.

(Loading Video...)

Movie 5.9. Daily average total number of infected nodes by closeness centralization quartile with random node infected first for connected graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.10. Daily average total number of infected nodes by closeness centralization quartile with most central node infected first for connected graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.11. Daily average total number of infected nodes by closeness centralization quartile and type of node infected first for connected graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.12. Daily average cumulative number of infected nodes by closeness centralization quartile with random node infected first for connected graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.13. Daily average cumulative number of infected nodes by closeness centralization quartile with most central node infected first for connected graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.14. Daily average cumulative number of infected nodes by closeness centralization quartile and type of node infected first for connected graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Table 5.11: Average epidemic duration by closeness centralization quartile for connected graphs above the epidemic threshold

| Closeness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Epidemic Duration | | | | | | |
| Connected Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| **Order** | **Cent Quartile** | **n** | **Mean** | **SD** | **Mean** | **SD** |
| 5 | Low | 628 | 12.55 | 0.77 | 12.65 | 0.72 |
| | Low-Mod | 1190 | 11.77 | 1.07 | 12.11 | 0.91 |
| | Mod-High | 963 | 11.59 | 1.10 | 12.01 | 0.86 |
| | High | 1469 | 11.21 | 0.90 | 11.79 | 0.71 |
| 10 | Low | 232 | 15.79 | 1.40 | 16.19 | 1.17 |
| | Low-Mod | 2031 | 15.18 | 1.35 | 16.01 | 1.05 |
| | Mod-High | 1435 | 14.90 | 1.34 | 15.88 | 0.91 |
| | High | 1859 | 13.53 | 1.21 | 15.03 | 0.82 |
| 15 | Low | 142 | 17.23 | 1.63 | 18.31 | 1.43 |
| | Low-Mod | 1791 | 17.37 | 1.46 | 18.66 | 1.00 |
| | Mod-High | 1534 | 16.72 | 1.44 | 18.17 | 0.93 |
| | High | 2163 | 14.83 | 1.29 | 16.90 | 0.85 |
| 20 | Low | 97 | 18.58 | 1.65 | 20.38 | 1.45 |
| | Low-Mod | 1299 | 18.78 | 1.47 | 20.53 | 0.96 |
| | Mod-High | 1176 | 18.36 | 1.47 | 19.96 | 0.85 |
| | High | 1788 | 15.49 | 1.23 | 18.04 | 0.76 |
| 25 | Low | 33 | 19.97 | 1.59 | 21.87 | 1.37 |
| | Low-Mod | 724 | 20.03 | 1.65 | 21.95 | 1.01 |
| | Mod-High | 792 | 19.38 | 1.51 | 21.29 | 0.90 |
| | High | 1758 | 16.18 | 1.42 | 19.05 | 0.84 |
| 30 | Low | 42 | 21.08 | 1.84 | 23.30 | 1.36 |
| | Low-Mod | 621 | 20.92 | 1.74 | 23.14 | 1.03 |
| | Mod-High | 656 | 20.00 | 1.66 | 22.21 | 0.97 |
| | High | 1464 | 16.54 | 1.25 | 19.77 | 0.77 |
| 35 | Low | 1 | 20.96 | NA | 26.83 | NA |
| | Low-Mod | 455 | 21.66 | 1.45 | 24.20 | 0.84 |
| | Mod-High | 547 | 21.03 | 1.61 | 23.19 | 0.98 |
| | High | 1477 | 16.91 | 1.25 | 20.41 | 0.73 |
| 40 | Low | 4 | 22.23 | 0.92 | 26.05 | 0.13 |
| | Low-Mod | 282 | 22.55 | 1.42 | 25.09 | 0.87 |
| | Mod-High | 332 | 22.10 | 1.64 | 24.18 | 0.93 |
| | High | 1393 | 17.26 | 1.22 | 20.96 | 0.73 |

Table 5.12: Parameter estimates from Cox proportional hazards model for epidemic duration for connected graphs above the epidemic threshold

| **Closeness Centralization** | | | | |
|---|---|---|---|---|
| **Epidemic Duration** | | | | |
| **Connected Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 4.0244 | 0.0285 | 55.9457 | $< 0.0001$ |
| Most Central Node First (ref: Random) | -0.3434 | 0.0256 | 0.7094 | $< 0.0001$ |
| Order of Largest Component | -0.1859 | 0.0008 | 0.8303 | $< 0.0001$ |
| Centralization*Most Central Node First | -0.9983 | 0.0356 | 0.3685 | $< 0.0001$ |

Table 5.13: Average peak number infected by closeness centralization quartile for connected graphs above the epidemic threshold

| Closeness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Peak Number Infected Connected Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 628 | 3.29 | 0.32 | 3.35 | 0.29 |
| | Low-Mod | 1190 | 2.87 | 0.42 | 3.07 | 0.38 |
| | Mod-High | 963 | 2.79 | 0.38 | 3.05 | 0.33 |
| | High | 1469 | 2.70 | 0.21 | 3.07 | 0.17 |
| 10 | Low | 232 | 4.90 | 1.00 | 5.25 | 0.93 |
| | Low-Mod | 2031 | 4.58 | 0.85 | 5.21 | 0.76 |
| | Mod-High | 1435 | 4.66 | 0.76 | 5.47 | 0.62 |
| | High | 1859 | 4.36 | 0.46 | 5.44 | 0.34 |
| 15 | Low | 142 | 5.55 | 1.23 | 6.34 | 1.16 |
| | Low-Mod | 1791 | 6.09 | 1.07 | 7.20 | 0.96 |
| | Mod-High | 1534 | 6.21 | 0.94 | 7.58 | 0.77 |
| | High | 2163 | 5.98 | 0.56 | 7.77 | 0.41 |
| 20 | Low | 97 | 6.24 | 1.27 | 7.51 | 1.26 |
| | Low-Mod | 1299 | 7.51 | 1.36 | 9.14 | 1.22 |
| | Mod-High | 1176 | 8.10 | 1.18 | 9.94 | 0.94 |
| | High | 1788 | 7.52 | 0.62 | 10.03 | 0.42 |
| 25 | Low | 33 | 6.88 | 0.81 | 8.66 | 0.75 |
| | Low-Mod | 724 | 8.84 | 1.82 | 10.90 | 1.61 |
| | Mod-High | 792 | 9.63 | 1.33 | 12.02 | 1.08 |
| | High | 1758 | 9.21 | 0.91 | 12.43 | 0.60 |
| 30 | Low | 42 | 9.05 | 2.79 | 10.90 | 2.58 |
| | Low-Mod | 621 | 10.20 | 1.94 | 12.72 | 1.70 |
| | Mod-High | 656 | 10.86 | 1.42 | 13.78 | 1.11 |
| | High | 1464 | 10.71 | 0.85 | 14.70 | 0.53 |
| 35 | Low | 1 | 8.71 | NA | 11.09 | NA |
| | Low-Mod | 455 | 11.38 | 1.86 | 14.45 | 1.59 |
| | Mod-High | 547 | 12.47 | 1.47 | 15.78 | 1.17 |
| | High | 1477 | 12.25 | 0.95 | 16.96 | 0.54 |
| 40 | Low | 4 | 10.85 | 0.56 | 13.82 | 0.42 |
| | Low-Mod | 282 | 13.38 | 2.13 | 16.84 | 1.90 |
| | Mod-High | 332 | 15.16 | 1.91 | 18.86 | 1.45 |
| | High | 1393 | 13.85 | 1.01 | 19.27 | 0.56 |

Table 5.14: Parameter estimates from linear regression model for peak number of infected nodes for connected graphs above the epidemic threshold

| **Closeness Centralization** | | | |
|---|---|---|---|
| **Peak Number Infected** | | | |
| **Connected Graphs Above Epidemic Threshold** | | | |
| **Parameter** | **Estimate** | **SE** | **p-value** |
| Intercept | 0.7122 | 0.0190 | $< 0.0001$ |
| Centralization | -0.7069 | 0.0255 | $< 0.0001$ |
| Most Central Node First (ref: Random) | -0.0013 | 0.0258 | 0.9604 |
| Order of Largest Component | 0.3714 | 0.0005 | $< 0.0001$ |
| Centralization*Most Central Node First | 2.9493 | 0.0353 | $< 0.0001$ |

Table 5.15: Average day peak number infected by closeness centralization quartile for connected graphs above the epidemic threshold

| Closeness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Day Peak Number Infected Connected Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 628 | 4.67 | 0.24 | 4.59 | 0.22 |
| | Low-Mod | 1190 | 4.62 | 0.24 | 4.44 | 0.20 |
| | Mod-High | 963 | 4.60 | 0.25 | 4.28 | 0.24 |
| | High | 1469 | 4.51 | 0.26 | 4.03 | 0.23 |
| 10 | Low | 232 | 6.09 | 0.39 | 5.90 | 0.33 |
| | Low-Mod | 2031 | 5.92 | 0.37 | 5.50 | 0.32 |
| | Mod-High | 1435 | 5.74 | 0.35 | 5.06 | 0.28 |
| | High | 1859 | 5.16 | 0.38 | 4.43 | 0.25 |
| 15 | Low | 142 | 6.94 | 0.52 | 6.63 | 0.48 |
| | Low-Mod | 1791 | 6.75 | 0.46 | 6.13 | 0.40 |
| | Mod-High | 1534 | 6.31 | 0.45 | 5.43 | 0.36 |
| | High | 2163 | 5.39 | 0.43 | 4.59 | 0.26 |
| 20 | Low | 97 | 7.61 | 0.57 | 7.27 | 0.54 |
| | Low-Mod | 1299 | 7.23 | 0.53 | 6.49 | 0.49 |
| | Mod-High | 1176 | 6.71 | 0.48 | 5.69 | 0.37 |
| | High | 1788 | 5.38 | 0.40 | 4.58 | 0.22 |
| 25 | Low | 33 | 8.33 | 0.61 | 7.96 | 0.68 |
| | Low-Mod | 724 | 7.67 | 0.59 | 6.75 | 0.52 |
| | Mod-High | 792 | 6.96 | 0.53 | 5.88 | 0.43 |
| | High | 1758 | 5.42 | 0.43 | 4.61 | 0.23 |
| 30 | Low | 42 | 8.48 | 0.51 | 8.16 | 0.55 |
| | Low-Mod | 621 | 7.92 | 0.62 | 6.91 | 0.60 |
| | Mod-High | 656 | 7.09 | 0.60 | 5.90 | 0.46 |
| | High | 1464 | 5.38 | 0.37 | 4.60 | 0.17 |
| 35 | Low | 1 | 8.15 | NA | 8.06 | NA |
| | Low-Mod | 455 | 8.10 | 0.54 | 7.06 | 0.48 |
| | Mod-High | 547 | 7.36 | 0.62 | 6.11 | 0.50 |
| | High | 1477 | 5.38 | 0.35 | 4.60 | 0.16 |
| 40 | Low | 4 | 8.76 | 0.31 | 7.99 | 0.35 |
| | Low-Mod | 282 | 8.29 | 0.63 | 7.12 | 0.57 |
| | Mod-High | 332 | 7.50 | 0.57 | 6.21 | 0.44 |
| | High | 1393 | 5.38 | 0.32 | 4.61 | 0.15 |

Table 5.16: Parameter estimates from Cox proportional hazards model for day peak number infected for connected graphs above the epidemic threshold

| **Closeness Centralization** | | | | |
|---|---|---|---|---|
| **Day Peak Number Infected** | | | | |
| **Connected Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 5.2914 | 0.0330 | 198.6206 | $< 0.0001$ |
| Most Central Node First (ref: Random) | 0.4805 | 0.0274 | 1.6169 | $< 0.0001$ |
| Order of Largest Component | -0.0960 | 0.0006 | 0.9084 | $< 0.0001$ |
| Centralization*Most Central Node First | 1.9259 | 0.0387 | 6.8611 | $< 0.0001$ |

Table 5.17: Average final cumulative number infected by closeness centralization quartile for connected graphs above the epidemic threshold

| Closeness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Final Cumulative Number Infected Connected Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 628 | 4.35 | 0.32 | 4.42 | 0.27 |
| | Low-Mod | 1190 | 3.83 | 0.54 | 4.04 | 0.47 |
| | Mod-High | 963 | 3.70 | 0.53 | 3.99 | 0.43 |
| | High | 1469 | 3.52 | 0.32 | 3.93 | 0.23 |
| 10 | Low | 232 | 7.65 | 1.36 | 8.11 | 1.20 |
| | Low-Mod | 2031 | 7.09 | 1.20 | 7.91 | 1.01 |
| | Mod-High | 1435 | 7.02 | 1.11 | 8.04 | 0.84 |
| | High | 1859 | 6.15 | 0.77 | 7.52 | 0.53 |
| 15 | Low | 142 | 9.68 | 1.98 | 10.94 | 1.82 |
| | Low-Mod | 1791 | 10.27 | 1.67 | 11.90 | 1.37 |
| | Mod-High | 1534 | 9.98 | 1.50 | 11.89 | 1.14 |
| | High | 2163 | 8.68 | 1.02 | 11.09 | 0.73 |
| 20 | Low | 97 | 11.77 | 2.31 | 14.02 | 2.26 |
| | Low-Mod | 1299 | 13.29 | 2.13 | 15.86 | 1.71 |
| | Mod-High | 1176 | 13.48 | 1.88 | 16.19 | 1.36 |
| | High | 1788 | 10.92 | 1.12 | 14.42 | 0.77 |
| 25 | Low | 33 | 13.94 | 1.77 | 17.12 | 1.79 |
| | Low-Mod | 724 | 16.27 | 2.93 | 19.70 | 2.33 |
| | Mod-High | 792 | 16.49 | 2.23 | 20.19 | 1.66 |
| | High | 1758 | 13.46 | 1.67 | 17.99 | 1.19 |
| 30 | Low | 42 | 18.14 | 4.54 | 21.69 | 4.03 |
| | Low-Mod | 621 | 19.28 | 3.34 | 23.73 | 2.63 |
| | Mod-High | 656 | 19.08 | 2.62 | 23.79 | 1.94 |
| | High | 1464 | 15.62 | 1.53 | 21.28 | 1.08 |
| 35 | Low | 1 | 19.09 | NA | 25.95 | NA |
| | Low-Mod | 455 | 21.97 | 2.95 | 27.60 | 2.15 |
| | Mod-High | 547 | 22.38 | 2.69 | 27.89 | 2.01 |
| | High | 1477 | 17.89 | 1.66 | 24.61 | 1.09 |
| 40 | Low | 4 | 22.87 | 0.99 | 29.91 | 0.48 |
| | Low-Mod | 282 | 25.97 | 3.35 | 32.36 | 2.36 |
| | Mod-High | 332 | 27.25 | 3.38 | 33.43 | 2.37 |
| | High | 1393 | 20.23 | 1.75 | 27.98 | 1.17 |

Table 5.18: Parameter estimates from linear regression model for final cumulative number of infected nodes for connected graphs above the epidemic threshold

| **Closeness Centralization** | | | |
|---|---|---|---|
| **Final Cumulative Number Infected Connected Graphs Above Epidemic Threshold** | | | |
| **Parameter** | **Estimate** | **SE** | **p-value** |
| Intercept | 2.6783 | 0.0321 | $< 0.0001$ |
| Centralization | -5.2036 | 0.0431 | $< 0.0001$ |
| Most Central Node First (ref: Random) | 0.4136 | 0.0436 | $< 0.0001$ |
| Order of Largest Component | 0.6498 | 0.0008 | $< 0.0001$ |
| Centralization*Most Central Node First | 3.7002 | 0.0597 | $< 0.0001$ |

Table 5.19: Average day final cumulative number infected by closeness centralization quartile for connected graphs above the epidemic threshold

| Closeness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Day Final Cumulative Number Infected Connected Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 628 | 5.82 | 0.45 | 5.77 | 0.45 |
| | Low-Mod | 1190 | 5.87 | 0.33 | 5.76 | 0.28 |
| | Mod-High | 963 | 5.81 | 0.33 | 5.57 | 0.28 |
| | High | 1469 | 5.56 | 0.39 | 5.18 | 0.32 |
| 10 | Low | 232 | 8.38 | 0.64 | 8.35 | 0.66 |
| | Low-Mod | 2031 | 8.10 | 0.57 | 8.02 | 0.50 |
| | Mod-High | 1435 | 7.74 | 0.54 | 7.47 | 0.45 |
| | High | 1859 | 6.58 | 0.67 | 6.26 | 0.58 |
| 15 | Low | 142 | 10.03 | 0.87 | 10.24 | 0.78 |
| | Low-Mod | 1791 | 9.67 | 0.71 | 9.68 | 0.60 |
| | Mod-High | 1534 | 8.95 | 0.73 | 8.80 | 0.56 |
| | High | 2163 | 7.08 | 0.88 | 6.85 | 0.81 |
| 20 | Low | 97 | 11.34 | 0.85 | 11.93 | 0.74 |
| | Low-Mod | 1299 | 10.72 | 0.83 | 10.93 | 0.74 |
| | Mod-High | 1176 | 9.91 | 0.78 | 9.78 | 0.65 |
| | High | 1788 | 7.09 | 0.90 | 6.93 | 0.87 |
| 25 | Low | 33 | 12.51 | 1.03 | 13.08 | 0.99 |
| | Low-Mod | 724 | 11.65 | 0.96 | 11.90 | 0.85 |
| | Mod-High | 792 | 10.61 | 0.86 | 10.61 | 0.70 |
| | High | 1758 | 7.22 | 1.03 | 7.09 | 0.97 |
| 30 | Low | 42 | 12.95 | 0.83 | 13.89 | 0.97 |
| | Low-Mod | 621 | 12.32 | 0.99 | 12.68 | 0.84 |
| | Mod-High | 656 | 11.18 | 1.01 | 11.31 | 0.77 |
| | High | 1464 | 7.14 | 0.95 | 7.04 | 0.94 |
| 35 | Low | 1 | 13.63 | NA | 17.04 | NA |
| | Low-Mod | 455 | 12.90 | 0.93 | 13.47 | 0.90 |
| | Mod-High | 547 | 11.89 | 1.02 | 12.00 | 0.85 |
| | High | 1477 | 7.16 | 0.93 | 7.11 | 0.94 |
| 40 | Low | 4 | 14.09 | 0.83 | 15.99 | 0.16 |
| | Low-Mod | 282 | 13.47 | 1.00 | 13.94 | 1.03 |
| | Mod-High | 332 | 12.31 | 0.94 | 12.33 | 0.75 |
| | High | 1393 | 7.14 | 0.89 | 7.07 | 0.93 |

Table 5.20: Parameter estimates from Cox proportional hazards model for day final cumulative number infected for connected graphs above the epidemic threshold

| **Closeness Centralization** | | | | |
|---|---|---|---|---|
| **Day Final Cumulative Number Infected Connected Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 5.2424 | 0.0317 | 189.1300 | $< 0.0001$ |
| Most Central Node First (ref: Random) | -0.3494 | 0.0256 | 0.7051 | $< 0.0001$ |
| Order of Largest Component | -0.1323 | 0.0007 | 0.8761 | $< 0.0001$ |
| Centralization*Most Central Node First | 0.7635 | 0.0357 | 2.1458 | $< 0.0001$ |

### 5.6.3 Betweenness Centralization

| Betweenness Centralization | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Graph Order (# nodes) | Number of Graphs | | | | | | | | |
| | Below Threshold | | | | Above Threshold | | | | |
| | Low | Low-Mod | Mod-High | High | Low | Low-Mod | Mod-High | High | Total |
| 5 | 193 | 10 | 0 | 0 | 2,526 | 1,882 | 937 | 1,016 | 6,564 |
| 10 | 185 | 6 | 0 | 0 | 3,930 | 5,243 | 3,305 | 2,101 | 14,770 |
| 15 | 80 | 0 | 0 | 0 | 4,057 | 7,882 | 6,276 | 4,113 | 22,408 |
| 20 | 39 | 0 | 0 | 0 | 3,858 | 10,834 | 9,809 | 5,460 | 30,000 |
| 25 | 31 | 0 | 0 | 0 | 3,772 | 13,396 | 13,055 | 7,238 | 37,492 |
| 30 | 11 | 0 | 0 | 0 | 3,260 | 14,686 | 18,208 | 8,835 | 45,000 |
| 35 | 17 | 0 | 0 | 0 | 2,852 | 15,009 | 24,004 | 10,618 | 52,500 |
| 40 | 0 | 0 | 0 | 0 | 2,102 | 16,601 | 29,031 | 12,266 | 60,000 |

Figure 5.3: Number of graphs generated by betweenness centralization category

Movie 5.15. Distribution of betweenness centralization values produced by the modified Star Start program for graphs of order 5-40 nodes.

(Loading Video...)

Movie 5.16. Daily average total number of infected nodes in the largest component by betweenness centralization quartile with random node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.17. Daily average total number of infected nodes in the largest component by betweenness centralization quartile with most central node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.18. Daily average total number of infected nodes in the largest component by betweenness centralization quartile and type of node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.19. Daily average cumulative number of infected nodes in the largest component by betweenness centralization quartile with random node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.20. Daily average cumulative number of infected nodes in the largest component by betweenness centralization quartile with most central node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.21. Daily average cumulative number of infected nodes in the largest component by betweenness centralization quartile and type of node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Table 5.21: Average epidemic duration by betweenness centralization quartile and largest component order for graphs above the epidemic threshold

| Betweenness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Epidemic Duration | | | | | | |
| Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 1999 | 11.66 | 1.57 | 12.09 | 1.10 |
| | Low-Mod | 1147 | 11.67 | 1.02 | 11.97 | 0.84 |
| | Mod-High | 937 | 11.32 | 0.94 | 11.84 | 0.76 |
| | High | 1016 | 10.90 | 0.80 | 11.58 | 0.67 |
| 10 | Low | 1630 | 14.46 | 1.98 | 15.44 | 1.44 |
| | Low-Mod | 2376 | 14.85 | 1.47 | 15.78 | 1.07 |
| | Mod-High | 1217 | 14.52 | 1.22 | 15.60 | 0.91 |
| | High | 1840 | 13.33 | 1.05 | 14.89 | 0.72 |
| 15 | Low | 1263 | 16.08 | 2.15 | 17.68 | 1.46 |
| | Low-Mod | 3385 | 16.27 | 1.80 | 17.89 | 1.20 |
| | Mod-High | 1616 | 16.32 | 1.51 | 17.92 | 1.02 |
| | High | 2488 | 14.89 | 1.27 | 16.94 | 0.83 |
| 20 | Low | 922 | 17.20 | 2.08 | 19.37 | 1.53 |
| | Low-Mod | 3834 | 17.38 | 1.89 | 19.53 | 1.23 |
| | Mod-High | 2812 | 17.01 | 1.76 | 19.14 | 1.11 |
| | High | 2107 | 15.73 | 1.38 | 18.21 | 0.87 |
| 25 | Low | 416 | 18.36 | 2.15 | 21.00 | 1.55 |
| | Low-Mod | 4153 | 18.07 | 1.98 | 20.69 | 1.29 |
| | Mod-High | 3916 | 17.68 | 1.75 | 20.17 | 1.13 |
| | High | 2053 | 16.47 | 1.56 | 19.27 | 1.01 |
| 30 | Low | 162 | 20.56 | 1.96 | 23.13 | 1.57 |
| | Low-Mod | 3308 | 19.28 | 1.97 | 22.14 | 1.25 |
| | Mod-High | 5928 | 18.26 | 1.81 | 21.07 | 1.20 |
| | High | 1838 | 17.05 | 1.66 | 20.15 | 1.09 |
| 35 | Low | 74 | 22.21 | 1.62 | 24.86 | 1.09 |
| | Low-Mod | 1219 | 21.06 | 1.75 | 23.99 | 1.01 |
| | Mod-High | 6183 | 19.30 | 1.89 | 22.23 | 1.29 |
| | High | 2044 | 17.34 | 1.71 | 20.72 | 1.07 |
| 40 | Low | 0 | NA | NA | NA | NA |
| | Low-Mod | 37 | 23.12 | 1.33 | 25.67 | 0.97 |
| | Mod-High | 474 | 22.61 | 1.33 | 24.78 | 0.80 |
| | High | 1500 | 17.51 | 1.52 | 21.14 | 0.98 |

Table 5.22: Parameter estimates from Cox proportional hazards model for epidemic duration for graphs above the epidemic threshold

| **Betweenness Centralization** | | | | |
|---|---|---|---|---|
| **Epidemic Duration** | | | | |
| **Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 2.7125 | 0.0106 | 15.0663 | $< 0.0001$ |
| Most Central Node First (ref: Random) | -0.7890 | 0.0083 | 0.4543 | $< 0.0001$ |
| Order of Largest Component | -0.1820 | 0.0003 | 0.8336 | $< 0.0001$ |
| Centralization*Most Central Node First | -0.7782 | 0.0143 | 0.4592 | $< 0.0001$ |

| Distribution of Spectral Radius | | | | |
|---|---|---|---|---|
| Order | n | 1Q | Median | 3Q |
| 5 | 6,361 | 1.73 | 2.17 | 2.69 |
| 10 | 14,579 | 2.55 | 3.00 | 3.33 |
| 15 | 22,328 | 3.08 | 3.53 | 3.84 |
| 20 | 29,961 | 3.42 | 3.93 | 4.31 |
| 25 | 37,461 | 3.69 | 4.24 | 4.71 |
| 30 | 44,989 | 4.01 | 4.59 | 5.07 |
| 35 | 52,483 | 4.28 | 4.91 | 5.41 |
| 40 | 60,000 | 4.56 | 5.16 | 5.75 |

Figure 5.4: Spectral radius of graphs above the epidemic threshold by graph order

Table 5.23: Average day of peak number of infected nodes by betweenness centralization quartile and largest component order for graphs above the epidemic threshold

| Betweenness Centralization | | | | | |
|---|---|---|---|---|---|
| | | Peak Number Infected | | | |
| | | Graphs Above Epidemic Threshold | | | |
| | | Random First | | Most Central First | |
| Order | Cent Quartile | Mean | SD | Mean | SD |
| 5 | Low | 2.90 | 0.57 | 3.10 | 0.45 |
| | Low-Mod | 2.78 | 0.34 | 3.00 | 0.33 |
| | Mod-High | 2.67 | 0.29 | 2.98 | 0.27 |
| | High | 2.60 | 0.16 | 3.00 | 0.14 |
| 10 | Low | 4.26 | 1.12 | 4.87 | 1.00 |
| | Low-Mod | 4.50 | 0.81 | 5.21 | 0.75 |
| | Mod-High | 4.35 | 0.68 | 5.18 | 0.65 |
| | High | 4.23 | 0.38 | 5.33 | 0.32 |
| 15 | Low | 5.19 | 1.29 | 6.33 | 1.19 |
| | Low-Mod | 5.69 | 1.12 | 7.01 | 1.02 |
| | Mod-High | 5.92 | 0.87 | 7.34 | 0.81 |
| | High | 5.89 | 0.57 | 7.66 | 0.48 |
| 20 | Low | 5.89 | 1.41 | 7.50 | 1.47 |
| | Low-Mod | 6.73 | 1.44 | 8.59 | 1.35 |
| | Mod-High | 7.35 | 1.09 | 9.45 | 0.94 |
| | High | 7.46 | 0.70 | 9.92 | 0.57 |
| 25 | Low | 6.47 | 1.62 | 8.38 | 1.65 |
| | Low-Mod | 7.77 | 1.65 | 10.23 | 1.62 |
| | Mod-High | 8.74 | 1.30 | 11.49 | 1.20 |
| | High | 9.11 | 0.88 | 12.25 | 0.70 |
| 30 | Low | 8.19 | 2.02 | 10.18 | 2.04 |
| | Low-Mod | 8.81 | 1.80 | 11.68 | 1.75 |
| | Mod-High | 10.13 | 1.41 | 13.50 | 1.32 |
| | High | 10.57 | 1.01 | 14.39 | 0.93 |
| 35 | Low | 9.57 | 1.66 | 11.98 | 1.96 |
| | Low-Mod | 10.17 | 2.10 | 13.27 | 2.04 |
| | Mod-High | 11.52 | 1.70 | 15.34 | 1.63 |
| | High | 12.16 | 0.99 | 16.74 | 0.84 |
| 40 | Low | NA | NA | NA | NA |
| | Low-Mod | 13.67 | 2.59 | 16.72 | 2.61 |
| | Mod-High | 14.57 | 2.26 | 18.06 | 2.02 |
| | High | 13.82 | 1.06 | 19.16 | 0.75 |

Table 5.24: Parameter estimates from linear regression model for peak number of infected nodes for graphs above the epidemic threshold

| **Betweenness Centralization** | | | |
|---|---|---|---|
| **Peak Number of Infected Nodes Graphs Above Epidemic Threshold** | | | |
| **Parameter** | **Estimate** | **SE** | **p-value** |
| Intercept | -0.4946 | 0.0079 | $< 0.0001$ |
| Centralization | 1.5727 | 0.0129 | $< 0.0001$ |
| Most Central Node First (ref: Random) | 0.7065 | 0.0101 | $< 0.0001$ |
| Order of Largest Component | 0.3230 | 0.0002 | $< 0.0001$ |
| Centralization*Most Central Node First | 3.3413 | 0.0171 | $< 0.0001$ |

Table 5.25: Average day of peak number of infected nodes by betweenness centralization quartile and largest component order for graphs above the epidemic threshold

| Betweenness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Day Peak Number Infected Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 1999 | 4.53 | 0.46 | 4.36 | 0.34 |
| | Low-Mod | 1147 | 4.64 | 0.25 | 4.39 | 0.19 |
| | Mod-High | 937 | 4.58 | 0.25 | 4.19 | 0.19 |
| | High | 1016 | 4.45 | 0.26 | 3.94 | 0.19 |
| 10 | Low | 1630 | 5.72 | 0.57 | 5.36 | 0.53 |
| | Low-Mod | 2376 | 5.79 | 0.45 | 5.28 | 0.46 |
| | Mod-High | 1217 | 5.71 | 0.34 | 5.04 | 0.26 |
| | High | 1840 | 5.12 | 0.34 | 4.41 | 0.22 |
| 15 | Low | 1263 | 6.43 | 0.74 | 5.92 | 0.70 |
| | Low-Mod | 3385 | 6.30 | 0.64 | 5.60 | 0.64 |
| | Mod-High | 1616 | 6.22 | 0.59 | 5.39 | 0.50 |
| | High | 2488 | 5.45 | 0.45 | 4.64 | 0.29 |
| 20 | Low | 922 | 6.93 | 0.77 | 6.45 | 0.80 |
| | Low-Mod | 3834 | 6.71 | 0.72 | 5.91 | 0.68 |
| | Mod-High | 2812 | 6.24 | 0.71 | 5.30 | 0.60 |
| | High | 2107 | 5.52 | 0.52 | 4.69 | 0.35 |
| 25 | Low | 416 | 7.54 | 0.80 | 7.19 | 0.68 |
| | Low-Mod | 4153 | 6.86 | 0.84 | 5.99 | 0.79 |
| | Mod-High | 3916 | 6.29 | 0.76 | 5.31 | 0.63 |
| | High | 2053 | 5.58 | 0.59 | 4.74 | 0.39 |
| 30 | Low | 162 | 8.38 | 0.76 | 8.00 | 0.57 |
| | Low-Mod | 3308 | 7.34 | 0.83 | 6.41 | 0.78 |
| | Mod-High | 5928 | 6.35 | 0.81 | 5.32 | 0.66 |
| | High | 1838 | 5.66 | 0.69 | 4.80 | 0.47 |
| 35 | Low | 74 | 9.07 | 0.65 | 8.68 | 0.69 |
| | Low-Mod | 1219 | 8.12 | 0.65 | 7.20 | 0.55 |
| | Mod-High | 6183 | 6.66 | 0.90 | 5.59 | 0.76 |
| | High | 2044 | 5.60 | 0.68 | 4.76 | 0.45 |
| 40 | Low | 0 | NA | NA | NA | NA |
| | Low-Mod | 37 | 8.70 | 0.76 | 7.71 | 0.77 |
| | Mod-High | 474 | 8.00 | 0.56 | 6.75 | 0.52 |
| | High | 1500 | 5.50 | 0.54 | 4.70 | 0.35 |

Table 5.26: Parameter estimates from Cox proportional hazards model for day peak number infected for graphs above the epidemic threshold

| **Betweenness Centralization** | | | | |
|---|---|---|---|---|
| **Day Peak Number Infected** | | | | |
| **Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 3.5649 | 0.0110 | 35.3362 | $< 0.0001$ |
| Most Central Node First (ref: Random) | 0.2287 | 0.0085 | 1.2570 | $< 0.0001$ |
| Order of Largest Component | -0.1007 | 0.0002 | 0.9042 | $< 0.0001$ |
| Centralization*Most Central Node First | 2.0292 | 0.0147 | 7.6082 | $< 0.0001$ |

Table 5.27: Average final cumulative number infected nodes by betweenness centralization quartile and largest component order for graphs above the epidemic threshold

| Betweenness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Final Cumulative Number Infected Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 1999 | 3.80 | 0.78 | 4.05 | 0.60 |
| | Low-Mod | 1147 | 3.73 | 0.45 | 3.96 | 0.40 |
| | Mod-High | 937 | 3.54 | 0.39 | 3.88 | 0.33 |
| | High | 1016 | 3.35 | 0.23 | 3.81 | 0.17 |
| 10 | Low | 1630 | 6.49 | 1.69 | 7.34 | 1.43 |
| | Low-Mod | 2376 | 6.87 | 1.20 | 7.79 | 0.99 |
| | Mod-High | 1217 | 6.58 | 0.98 | 7.65 | 0.83 |
| | High | 1840 | 5.96 | 0.60 | 7.37 | 0.43 |
| 15 | Low | 1263 | 8.71 | 2.21 | 10.44 | 1.87 |
| | Low-Mod | 3385 | 9.26 | 1.84 | 11.18 | 1.50 |
| | Mod-High | 1616 | 9.47 | 1.41 | 11.49 | 1.12 |
| | High | 2488 | 8.63 | 0.97 | 11.03 | 0.71 |
| 20 | Low | 922 | 10.55 | 2.54 | 13.25 | 2.38 |
| | Low-Mod | 3834 | 11.53 | 2.40 | 14.46 | 2.01 |
| | Mod-High | 2812 | 11.84 | 1.89 | 14.95 | 1.41 |
| | High | 2107 | 11.04 | 1.23 | 14.48 | 0.86 |
| 25 | Low | 416 | 12.45 | 3.06 | 15.95 | 2.91 |
| | Low-Mod | 4153 | 13.62 | 2.79 | 17.66 | 2.34 |
| | Mod-High | 3916 | 14.21 | 2.16 | 18.39 | 1.65 |
| | High | 2053 | 13.59 | 1.63 | 18.07 | 1.17 |
| 30 | Low | 162 | 16.66 | 3.67 | 20.62 | 3.61 |
| | Low-Mod | 3308 | 16.27 | 3.17 | 21.32 | 2.67 |
| | Mod-High | 5928 | 16.66 | 2.36 | 21.91 | 1.76 |
| | High | 1838 | 15.96 | 1.85 | 21.51 | 1.34 |
| 35 | Low | 74 | 20.38 | 2.96 | 25.22 | 3.30 |
| | Low-Mod | 1219 | 20.05 | 3.58 | 25.87 | 3.13 |
| | Mod-High | 6183 | 19.63 | 2.75 | 25.83 | 2.03 |
| | High | 2044 | 18.24 | 2.00 | 24.88 | 1.34 |
| 40 | Low | 0 | NA | NA | NA | NA |
| | Low-Mod | 37 | 27.00 | 3.24 | 33.09 | 2.41 |
| | Mod-High | 474 | 27.25 | 3.27 | 33.39 | 2.28 |
| | High | 1500 | 20.48 | 2.05 | 28.18 | 1.43 |

Table 5.28: Parameter estimates from linear regression model for final cumulative number of infected nodes for graphs above the epidemic threshold

| **Betweenness Centralization** | | | |
|---|---|---|---|
| **Final Cumulative Number of Infected Nodes Graphs Above Epidemic Threshold** | | | |
| **Parameter** | **Estimate** | **SE** | **p-value** |
| Intercept | -0.7102 | 0.0126 | $< 0.0001$ |
| Centralization | -2.0482 | 0.0205 | $< 0.0001$ |
| Most Central Node First (ref: Random) | 1.5467 | 0.0161 | $< 0.0001$ |
| Order of Largest Component | 0.6367 | 0.0004 | $< 0.0001$ |
| Centralization*Most Central Node First | 4.4339 | 0.0273 | $< 0.0001$ |

Table 5.29: Average day final cumulative number of nodes infected by betweenness centralization quartile and largest component order for graphs above the epidemic threshold

| Betweenness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Day Final Cumulative Number Infected Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 1999 | 5.64 | 0.66 | 5.57 | 0.48 |
| | Low-Mod | 1147 | 5.94 | 0.33 | 5.72 | 0.28 |
| | Mod-High | 937 | 5.78 | 0.33 | 5.49 | 0.25 |
| | High | 1016 | 5.42 | 0.35 | 5.05 | 0.26 |
| 10 | Low | 1630 | 7.75 | 0.90 | 7.75 | 0.74 |
| | Low-Mod | 2376 | 7.86 | 0.74 | 7.70 | 0.71 |
| | Mod-High | 1217 | 7.72 | 0.51 | 7.49 | 0.41 |
| | High | 1840 | 6.52 | 0.61 | 6.23 | 0.55 |
| 15 | Low | 1263 | 9.14 | 1.22 | 9.35 | 0.99 |
| | Low-Mod | 3385 | 8.90 | 1.07 | 8.93 | 0.98 |
| | Mod-High | 1616 | 8.76 | 1.08 | 8.66 | 1.04 |
| | High | 2488 | 7.22 | 0.93 | 7.03 | 0.89 |
| 20 | Low | 922 | 10.17 | 1.26 | 10.73 | 1.06 |
| | Low-Mod | 3834 | 9.83 | 1.20 | 10.11 | 1.10 |
| | Mod-High | 2812 | 9.01 | 1.28 | 9.04 | 1.23 |
| | High | 2107 | 7.42 | 1.18 | 7.28 | 1.18 |
| 25 | Low | 416 | 11.32 | 1.34 | 12.23 | 0.98 |
| | Low-Mod | 4153 | 10.32 | 1.43 | 10.80 | 1.41 |
| | Mod-High | 3916 | 9.34 | 1.44 | 9.48 | 1.43 |
| | High | 2053 | 7.63 | 1.40 | 7.54 | 1.43 |
| 30 | Low | 162 | 12.84 | 1.23 | 13.83 | 0.94 |
| | Low-Mod | 3308 | 11.38 | 1.33 | 12.06 | 1.21 |
| | Mod-High | 5928 | 9.67 | 1.58 | 9.92 | 1.62 |
| | High | 1838 | 7.84 | 1.69 | 7.83 | 1.80 |
| 35 | Low | 74 | 14.19 | 1.05 | 15.19 | 0.80 |
| | Low-Mod | 1219 | 12.87 | 1.07 | 13.73 | 0.91 |
| | Mod-High | 6183 | 10.51 | 1.71 | 10.90 | 1.77 |
| | High | 2044 | 7.73 | 1.72 | 7.73 | 1.78 |
| 40 | Low | 0 | NA | NA | NA | NA |
| | Low-Mod | 37 | 13.94 | 1.35 | 14.50 | 1.63 |
| | Mod-High | 474 | 13.05 | 0.90 | 13.23 | 1.05 |
| | High | 1500 | 7.46 | 1.45 | 7.42 | 1.56 |

Table 5.30: Parameter estimates from Cox proportional hazards model for day last new node infected for graphs above the epidemic threshold

| **Betweenness Centralization** | | | | |
|---|---|---|---|---|
| **Day Final Cumulative Number Infected Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 3.9507 | 0.0112 | 51.9704 | $< 0.0001$ |
| Most Central Node First (ref: Random) | -0.4414 | 0.0087 | 0.6431 | $< 0.0001$ |
| Order of Largest Component | -0.1530 | 0.0002 | 0.8581 | $< 0.0001$ |
| Centralization*Most Central Node First | 0.6075 | 0.0150 | 1.8358 | $< 0.0001$ |

## 5.7 Appendix

### 5.7.1 Closeness

The results of the analysis of closeness centralization on all the network generated (connected and disconnected) are described in this section. The number of graphs in each closeness centralization quartile as well as the total number of graphs generated is shown in Figure 5.5. The figure also describes the number of graphs above and below the epidemic threshold. Again, the modified Star Start program produces very few graphs below the epidemic threshold, especially as graph order increases. Almost all the graphs below the epidemic threshold are in the low centralization quartile. No graphs below the epidemic threshold are produced when the graph order is 40. Movie 5.22 illustrates the actual distribution of closeness centralization values by graph order, regardless of epidemic threshold. The majority of graphs generated are lowly centralized, with relatively fewer higher centralized graphs as graph order increases.

Movie 5.23 plots the daily average total number of infected nodes in the largest component by closeness centralization quartile and epidemic threshold for a random node infected first. Movie 5.24 plots the daily average total number of infected nodes in the largest component by closeness centralization quartile and epidemic threshold for the most central node infected first. Movie 5.25 plots the daily average total number of infected nodes in the largest component by type of node infected first and epidemic threshold within a closeness centralization quartile. Similarly, Movies 5.26-5.28 plot the daily average cumulative number of infected nodes in the largest component by closeness centralization quartile and epidemic threshold for a random node infected first, for the most central node infected first, and then by type of node infected first and epidemic threshold within a closeness centralization quartile. As seen with degree centralization, graphs below the epidemic threshold have much shorter epidemics with many fewer nodes infected and there is no effect of centralization quartile. As shown in Movies 5.23 and 5.24, there is a very clear difference in the incidence curve between the low centralization quartile and the other quartiles, regardless of which node is infected first. The differences between the upper three quartiles are harder to

distinguish although a little clearer when a random node is infected first. The incidence curve has a higher peak and reaches the peak earlier as centralization quartile increases from low to moderate-high.

**Epidemic Duration** Table 5.31 describes the average epidemic duration for graphs that are above the epidemic threshold. As shown in Table 5.31, the duration of the epidemic lengthens as graph order increases and the epidemic is also longer when the most central node is infected first compared to a random node infected first. Epidemic duration also decreases as centralization increases from low-moderate to high levels. The results for graphs below the epidemic threshold are the same as described for degree and betweenness centralization. As shown in Table 5.32, all of the variables considered in the linear regression model are significantly associated ($p<0.0001$) with epidemic duration. As centralization increases, epidemic duration decreases although this effect is mitigated when the most central node is infected first. The adjusted R-squared for the model is 0.6802.

**Peak Number Infected** Table 5.33 describes the peak number of infected nodes for graphs that are above the epidemic threshold. As shown in Table 5.33, the peak number of infected nodes increases as closeness centralization quartile increases when the most central node is infected first. When a random node is infected first, the peak number of infected nodes does increase as centralization increases from low to moderate-high and then decreases when centralization is high. As graph order increases, the peak number of infected nodes also increases. The results for graphs below the epidemic threshold are the same as described for degree and betweenness centralization. As shown in Table 5.34, all of the variables in the linear regression model are significantly associated ($p<0.0001$) with number of nodes infected at the peak of the epidemic. Specifically, as centralization increases the peak number of infected nodes increases. The peak number of infected nodes is also larger if the most central node is infected first. The adjusted R-squared of the model is 0.8389.

**Peak Day** Table 5.35 describes the average day of the peak number of infected nodes for graphs that are above the epidemic threshold. As shown in Table 5.35, the day of the peak occurs earlier when the most central node is infected first, regardless of centralization quartile and largest component order. For both random node and most central node, the day of the peak occurs later as centralization increases from low to low-moderate levels and then changes to come earlier as centralization increases to high levels. As largest component order increases, the day of the peak occurs later. The results for graphs below the epidemic threshold are the same as described for degree and betweenness centralization. Parameter estimates and hazard ratios produced by a Cox proportional hazard model of peak day are shown in Table 5.36. As shown in the table, increasing centralization increases the risk of earlier time to peak number of infected nodes.

**Cumulative Number Infected** Table 5.37 describes the average final cumulative number of infected nodes for graphs that are above the epidemic threshold. As shown in Table 5.37, when the most central node is infected first the final cumulative number of infected nodes increases as centralization quartile increases from low to moderate-high although there is only a very small difference between low-moderate and moderate-high quartiles. Interestingly, graphs in the highest closeness centralization quartile have a smaller number of nodes infected than the graphs in the low-moderate quartile. A similar pattern is produced when a random node is infected first, although the difference between low and low-moderate quartiles is much larger. At all centralization levels and graph orders, infecting the most central node first produces a larger cumulative number of infected nodes than when a random node is infected first. Final cumulative number of infected nodes increases as largest component order increases regardless of centralization level or which node was infected first. The results for graphs below the epidemic threshold are the same as described for degree and betweenness centralization. As shown in Table 5.38, all of the variables in the linear regression model are significantly associated ($p<0.0001$) with final cumulative number of infected nodes. The adjusted R-squared of the model is 0.8778. The final cumulative number of nodes infected is much larger when the most central node is infected first.

**Day Final Number Infected** Table 5.39 describes the day that the final cumulative number of infected nodes was reached for graphs that are above the epidemic threshold. As shown in Table 5.39, the day of the last new infection occurs later as centralization quartile increases from low to low-moderate and the occurs earlier as centralization increases to high levels, regardless of which node is infected first. There is a very large difference in the day that the final new node is infected between high closeness centralization quartile and all of the other quartiles. The day when no new nodes are infected increases slightly as graph order increases. There is very little difference between random and most central node infected first. The results for graphs below the epidemic threshold are the same as described for degree and betweenness centralization. As shown in Table 5.40, all of the variables in the Cox proportional hazard model are significantly associated ($p<0.0001$) with day of last new infection. As centralization increases, the risk of a shorter time to day of last new infection increases.

### 5.7.2 Figures, Movies, Tables

**Closeness**

| Closeness Centralization | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Graph Order (# nodes) | Number of Graphs | | | | | | | | |
| | Below Threshold | | | | Above Threshold | | | | |
| | Low | Low-Mod | Mod-High | High | Low | Low-Mod | Mod-High | High | Total |
| 5 | 193 | 10 | 0 | 0 | 2,004 | 1,925 | 963 | 1,469 | 6,564 |
| 10 | 191 | 0 | 0 | 0 | 7,043 | 4,242 | 1,435 | 1,859 | 14,770 |
| 15 | 80 | 0 | 0 | 0 | 13,459 | 5,172 | 1,534 | 2,163 | 22,408 |
| 20 | 39 | 0 | 0 | 0 | 22,021 | 4,976 | 1,176 | 1,788 | 30,000 |
| 25 | 31 | 0 | 0 | 0 | 30,776 | 4,135 | 792 | 1,758 | 37,492 |
| 30 | 11 | 0 | 0 | 0 | 38,949 | 3,920 | 656 | 1,464 | 45,000 |
| 35 | 17 | 0 | 0 | 0 | 46,927 | 3,532 | 547 | 1,477 | 52,500 |
| 40 | 0 | 0 | 0 | 0 | 55,409 | 2,866 | 332 | 1,393 | 60,000 |

Figure 5.5: Number of graphs generated by closeness centralization category

Movie 5.22. Distribution of closeness centralization values produced by the modified Star Start program for graphs of order 5-40 nodes.

(Loading Video...)

Movie 5.23. Daily average total number of infected nodes in the largest component by closeness centralization quartile with random node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.24. Daily average total number of infected nodes in the largest component by closeness centralization quartile with most central node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.25. Daily average total number of infected nodes in the largest component by closeness centralization quartile and type of node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.26. Daily average cumulative number of infected nodes in the largest component by closeness centralization quartile with random node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.27. Daily average cumulative number of infected nodes in the largest component by closeness centralization quartile with most central node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Movie 5.28. Daily average cumulative number of infected nodes in the largest component by closeness centralization quartile and type of node infected first for graphs of order 5-40, probability of transmission 0.3 and probability of recovery 0.2.

(Loading Video...)

Table 5.31: Average epidemic duration by closeness centralization quartile and largest component order for graphs above the epidemic threshold

| Closeness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| | | Epidemic Duration | | | | |
| | | Graphs Above Epidemic Threshold | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 1477 | 11.34 | 1.67 | 11.83 | 1.13 |
| | Low-Mod | 1190 | 11.77 | 1.07 | 12.11 | 0.91 |
| | Mod-High | 963 | 11.59 | 1.10 | 12.01 | 0.86 |
| | High | 1469 | 11.21 | 0.90 | 11.79 | 0.71 |
| 10 | Low | 1738 | 13.61 | 1.68 | 14.88 | 1.20 |
| | Low-Mod | 2031 | 15.18 | 1.35 | 16.01 | 1.05 |
| | Mod-High | 1435 | 14.90 | 1.34 | 15.88 | 0.91 |
| | High | 1859 | 13.53 | 1.21 | 15.03 | 0.82 |
| 15 | Low | 3264 | 15.31 | 1.63 | 17.20 | 1.08 |
| | Low-Mod | 1791 | 17.37 | 1.46 | 18.66 | 1.00 |
| | Mod-High | 1534 | 16.72 | 1.44 | 18.17 | 0.93 |
| | High | 2163 | 14.83 | 1.29 | 16.90 | 0.85 |
| 20 | Low | 5412 | 16.59 | 1.67 | 18.95 | 1.13 |
| | Low-Mod | 1299 | 18.78 | 1.47 | 20.53 | 0.96 |
| | Mod-High | 1176 | 18.36 | 1.47 | 19.96 | 0.85 |
| | High | 1788 | 15.49 | 1.23 | 18.04 | 0.76 |
| 25 | Low | 7264 | 17.54 | 1.72 | 20.24 | 1.17 |
| | Low-Mod | 724 | 20.03 | 1.65 | 21.95 | 1.01 |
| | Mod-High | 792 | 19.38 | 1.51 | 21.29 | 0.90 |
| | High | 1758 | 16.18 | 1.42 | 19.05 | 0.84 |
| 30 | Low | 8495 | 18.41 | 1.82 | 21.31 | 1.26 |
| | Low-Mod | 621 | 20.92 | 1.74 | 23.14 | 1.03 |
| | Mod-High | 656 | 20.00 | 1.66 | 22.21 | 0.97 |
| | High | 1464 | 16.54 | 1.25 | 19.77 | 0.77 |
| 35 | Low | 7041 | 19.28 | 1.96 | 22.30 | 1.41 |
| | Low-Mod | 455 | 21.66 | 1.45 | 24.20 | 0.84 |
| | Mod-High | 547 | 21.03 | 1.61 | 23.19 | 0.98 |
| | High | 1477 | 16.91 | 1.25 | 20.41 | 0.73 |
| 40 | Low | 4 | 22.23 | 0.92 | 26.05 | 0.13 |
| | Low-Mod | 282 | 22.55 | 1.42 | 25.09 | 0.87 |
| | Mod-High | 332 | 22.10 | 1.64 | 24.18 | 0.93 |
| | High | 1393 | 17.26 | 1.22 | 20.96 | 0.73 |

Table 5.32: Parameter estimates from Cox proportional hazards model for epidemic duration for graphs above the epidemic threshold

| **Closeness Centralization** | | | | |
|---|---|---|---|---|
| **Epidemic Duration** | | | | |
| **Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 0.0880 | 0.0113 | 1.0920 | $< 0.0001$ |
| Most Central Node First (ref: Random) | -1.1855 | 0.0040 | 0.3056 | $< 0.0001$ |
| Order of Largest Component | -0.1498 | 0.0002 | 0.8609 | $< 0.0001$ |
| Centralization*Most Central Node First | 0.5288 | 0.0150 | 1.6969 | $< 0.0001$ |

Table 5.33: Average day of peak number of infected nodes by closeness centralization quartile and largest component order for graphs above the epidemic threshold

| Closeness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| | | | Peak Number Infected | | | |
| | | | Graphs Above Epidemic Threshold | | | |
| | | | | Random First | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 1477 | 2.74 | 0.58 | 2.97 | 0.45 |
| | Low-Mod | 1190 | 2.87 | 0.42 | 3.07 | 0.38 |
| | Mod-High | 963 | 2.79 | 0.38 | 3.05 | 0.33 |
| | High | 1469 | 2.70 | 0.21 | 3.07 | 0.17 |
| 10 | Low | 1738 | 3.80 | 0.79 | 4.57 | 0.75 |
| | Low-Mod | 2031 | 4.58 | 0.85 | 5.21 | 0.76 |
| | Mod-High | 1435 | 4.66 | 0.76 | 5.47 | 0.62 |
| | High | 1859 | 4.36 | 0.46 | 5.44 | 0.34 |
| 15 | Low | 3264 | 5.11 | 0.93 | 6.54 | 1.00 |
| | Low-Mod | 1791 | 6.09 | 1.07 | 7.20 | 0.96 |
| | Mod-High | 1534 | 6.21 | 0.94 | 7.58 | 0.77 |
| | High | 2163 | 5.98 | 0.56 | 7.77 | 0.41 |
| 20 | Low | 5412 | 6.45 | 1.20 | 8.48 | 1.31 |
| | Low-Mod | 1299 | 7.51 | 1.36 | 9.14 | 1.22 |
| | Mod-High | 1176 | 8.10 | 1.18 | 9.94 | 0.94 |
| | High | 1788 | 7.52 | 0.62 | 10.03 | 0.42 |
| 25 | Low | 7264 | 7.94 | 1.49 | 10.58 | 1.61 |
| | Low-Mod | 724 | 8.84 | 1.82 | 10.90 | 1.61 |
| | Mod-High | 792 | 9.63 | 1.33 | 12.02 | 1.08 |
| | High | 1758 | 9.21 | 0.91 | 12.43 | 0.60 |
| 30 | Low | 8495 | 9.51 | 1.65 | 12.76 | 1.77 |
| | Low-Mod | 621 | 10.20 | 1.94 | 12.72 | 1.70 |
| | Mod-High | 656 | 10.86 | 1.42 | 13.78 | 1.11 |
| | High | 1464 | 10.71 | 0.85 | 14.70 | 0.53 |
| 35 | Low | 7041 | 11.24 | 1.80 | 15.04 | 1.91 |
| | Low-Mod | 455 | 11.38 | 1.86 | 14.45 | 1.59 |
| | Mod-High | 547 | 12.47 | 1.47 | 15.78 | 1.17 |
| | High | 1477 | 12.25 | 0.95 | 16.96 | 0.54 |
| 40 | Low | 4 | 10.85 | 0.56 | 13.82 | 0.42 |
| | Low-Mod | 282 | 13.38 | 2.13 | 16.84 | 1.90 |
| | Mod-High | 332 | 15.16 | 1.91 | 18.86 | 1.45 |
| | High | 1393 | 13.85 | 1.01 | 19.27 | 0.56 |

Table 5.34: Parameter estimates from linear regression model for peak number of infected nodes for graphs above the epidemic threshold

| **Closeness Centralization** | | | |
|---|---|---|---|
| **Peak Number of Infected Nodes Graphs Above Epidemic Threshold** | | | |
| **Parameter** | **Estimate** | **SE** | **p-value** |
| Intercept | -1.0234 | 0.0071 | $< 0.0001$ |
| Centralization | 2.0336 | 0.0138 | $< 0.0001$ |
| Most Central Node First (ref: Random) | 2.5799 | 0.0055 | $< 0.0001$ |
| Order of Largest Component | 0.3655 | 0.0002 | $< 0.0001$ |
| Centralization*Most Central Node First | -0.2068 | 0.0195 | $< 0.0001$ |

Table 5.35: Average day of peak number of infected nodes by closeness centralization quartile and largest component order for graphs above the epidemic threshold

| Closeness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Day Peak Number Infected Graphs Above Epidemic Threshold | | | | | | |
| Order | Cent Quartile | n | Random First | | Most Central First | |
| | | | Mean | SD | Mean | SD |
| 5 | Low | 1477 | 4.48 | 0.51 | 4.32 | 0.36 |
| | Low-Mod | 1190 | 4.62 | 0.24 | 4.44 | 0.20 |
| | Mod-High | 963 | 4.60 | 0.25 | 4.28 | 0.24 |
| | High | 1469 | 4.51 | 0.26 | 4.03 | 0.23 |
| 10 | Low | 1738 | 5.52 | 0.58 | 5.09 | 0.56 |
| | Low-Mod | 2031 | 5.92 | 0.37 | 5.50 | 0.32 |
| | Mod-High | 1435 | 5.74 | 0.35 | 5.06 | 0.28 |
| | High | 1859 | 5.16 | 0.38 | 4.43 | 0.25 |
| 15 | Low | 3264 | 6.02 | 0.68 | 5.35 | 0.67 |
| | Low-Mod | 1791 | 6.75 | 0.46 | 6.13 | 0.40 |
| | Mod-High | 1534 | 6.31 | 0.45 | 5.43 | 0.36 |
| | High | 2163 | 5.39 | 0.43 | 4.59 | 0.26 |
| 20 | Low | 5412 | 6.36 | 0.75 | 5.56 | 0.77 |
| | Low-Mod | 1299 | 7.23 | 0.53 | 6.49 | 0.49 |
| | Mod-High | 1176 | 6.71 | 0.48 | 5.69 | 0.37 |
| | High | 1788 | 5.38 | 0.40 | 4.58 | 0.22 |
| 25 | Low | 7264 | 6.49 | 0.83 | 5.61 | 0.83 |
| | Low-Mod | 724 | 7.67 | 0.59 | 6.75 | 0.52 |
| | Mod-High | 792 | 6.96 | 0.53 | 5.88 | 0.43 |
| | High | 1758 | 5.42 | 0.43 | 4.61 | 0.23 |
| 30 | Low | 8495 | 6.62 | 0.93 | 5.65 | 0.89 |
| | Low-Mod | 621 | 7.92 | 0.62 | 6.91 | 0.60 |
| | Mod-High | 656 | 7.09 | 0.60 | 5.90 | 0.46 |
| | High | 1464 | 5.38 | 0.37 | 4.60 | 0.17 |
| 35 | Low | 7041 | 6.75 | 1.04 | 5.73 | 0.97 |
| | Low-Mod | 455 | 8.10 | 0.54 | 7.06 | 0.48 |
| | Mod-High | 547 | 7.36 | 0.62 | 6.11 | 0.50 |
| | High | 1477 | 5.38 | 0.35 | 4.60 | 0.16 |
| 40 | Low | 4 | 8.76 | 0.31 | 7.99 | 0.35 |
| | Low-Mod | 282 | 8.29 | 0.63 | 7.12 | 0.57 |
| | Mod-High | 332 | 7.50 | 0.57 | 6.21 | 0.44 |
| | High | 1393 | 5.38 | 0.32 | 4.61 | 0.15 |

Table 5.36: Parameter estimates from Cox proportional hazards model for day peak number infected for graphs above the epidemic threshold

| **Closeness Centralization** | | | | |
|---|---|---|---|---|
| **Day Peak Number Infected** | | | | |
| **Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 1.0383 | 0.0101 | 2.8244 | $< 0.0001$ |
| Most Central Node First (ref: Random) | 0.7461 | 0.0039 | 2.1087 | $< 0.0001$ |
| Order of Largest Component | -0.0484 | 0.0002 | 0.9527 | $< 0.0001$ |
| Centralization*Most Central Node First | 0.6788 | 0.0145 | 1.9715 | $< 0.0001$ |

Table 5.37: Average final cumulative number infected nodes by closeness centralization quartile and largest component order for graphs above the epidemic threshold

| Closeness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| **Final Cumulative Number Infected Graphs Above Epidemic Threshold** | | | | | | |
| **Order** | **Cent Quartile** | **n** | **Random First** | | **Most Central First** | |
| | | | **Mean** | **SD** | **Mean** | **SD** |
| 5 | Low | 1477 | 3.60 | 0.80 | 3.88 | 0.61 |
| | Low-Mod | 1190 | 3.83 | 0.54 | 4.04 | 0.47 |
| | Mod-High | 963 | 3.70 | 0.53 | 3.99 | 0.43 |
| | High | 1469 | 3.52 | 0.32 | 3.93 | 0.23 |
| 10 | Low | 1738 | 5.73 | 1.23 | 6.81 | 1.06 |
| | Low-Mod | 2031 | 7.09 | 1.20 | 7.91 | 1.01 |
| | Mod-High | 1435 | 7.02 | 1.11 | 8.04 | 0.84 |
| | High | 1859 | 6.15 | 0.77 | 7.52 | 0.53 |
| 15 | Low | 3264 | 8.16 | 1.43 | 10.28 | 1.28 |
| | Low-Mod | 1791 | 10.27 | 1.67 | 11.90 | 1.37 |
| | Mod-High | 1534 | 9.98 | 1.50 | 11.89 | 1.14 |
| | High | 2163 | 8.68 | 1.02 | 11.09 | 0.73 |
| 20 | Low | 5412 | 10.68 | 1.80 | 13.83 | 1.65 |
| | Low-Mod | 1299 | 13.29 | 2.13 | 15.86 | 1.71 |
| | Mod-High | 1176 | 13.48 | 1.88 | 16.19 | 1.36 |
| | High | 1788 | 10.92 | 1.12 | 14.42 | 0.77 |
| 25 | Low | 7264 | 13.32 | 2.18 | 17.52 | 1.90 |
| | Low-Mod | 724 | 16.27 | 2.93 | 19.70 | 2.33 |
| | Mod-High | 792 | 16.49 | 2.23 | 20.19 | 1.66 |
| | High | 1758 | 13.46 | 1.67 | 17.99 | 1.19 |
| 30 | Low | 8495 | 16.15 | 2.41 | 21.40 | 1.99 |
| | Low-Mod | 621 | 19.28 | 3.34 | 23.73 | 2.63 |
| | Mod-High | 656 | 19.08 | 2.62 | 23.79 | 1.94 |
| | High | 1464 | 15.62 | 1.53 | 21.28 | 1.08 |
| 35 | Low | 7041 | 19.30 | 2.70 | 25.54 | 2.10 |
| | Low-Mod | 455 | 21.97 | 2.95 | 27.60 | 2.15 |
| | Mod-High | 547 | 22.38 | 2.69 | 27.89 | 2.01 |
| | High | 1477 | 17.89 | 1.66 | 24.61 | 1.09 |
| 40 | Low | 4 | 22.87 | 0.99 | 29.91 | 0.48 |
| | Low-Mod | 282 | 25.97 | 3.35 | 32.36 | 2.36 |
| | Mod-High | 332 | 27.25 | 3.38 | 33.43 | 2.37 |
| | High | 1393 | 20.23 | 1.75 | 27.98 | 1.17 |

Table 5.38: Parameter estimates from linear regression model for final cumulative number of infected nodes for graphs above the epidemic threshold

| **Closeness Centralization** | | | |
|---|---|---|---|
| **Final Cumulative Number of Infected Nodes Graphs Above Epidemic Threshold** | | | |
| **Parameter** | **Estimate** | **SE** | **p-value** |
| Intercept | -2.2487 | 0.0104 | $< 0.0001$ |
| Centralization | 1.5814 | 0.0205 | $< 0.0001$ |
| Most Central Node First (ref: Random) | 4.1574 | 0.0081 | $< 0.0001$ |
| Order of Largest Component | 0.6421 | 0.0003 | $< 0.0001$ |
| Centralization*Most Central Node First | -0.9767 | 0.0289 | $< 0.0001$ |

Table 5.39: Average day final cumulative number of nodes infected by closeness centralization quartile and largest component order for graphs above the epidemic threshold

| Closeness Centralization | | | | | | |
|---|---|---|---|---|---|---|
| Day Final Cumulative Number Infected Graphs Above Epidemic Threshold | | | | | | |
| | | | Random First | | Most Central First | |
| Order | Cent Quartile | n | Mean | SD | Mean | SD |
| 5 | Low | 1477 | 5.60 | 0.75 | 5.55 | 0.54 |
| | Low-Mod | 1190 | 5.87 | 0.33 | 5.76 | 0.28 |
| | Mod-High | 963 | 5.81 | 0.33 | 5.57 | 0.28 |
| | High | 1469 | 5.56 | 0.39 | 5.18 | 0.32 |
| 10 | Low | 1738 | 7.42 | 0.93 | 7.44 | 0.85 |
| | Low-Mod | 2031 | 8.10 | 0.57 | 8.02 | 0.50 |
| | Mod-High | 1435 | 7.74 | 0.54 | 7.47 | 0.45 |
| | High | 1859 | 6.58 | 0.67 | 6.26 | 0.58 |
| 15 | Low | 3264 | 8.41 | 1.16 | 8.54 | 1.13 |
| | Low-Mod | 1791 | 9.67 | 0.71 | 9.68 | 0.60 |
| | Mod-High | 1534 | 8.95 | 0.73 | 8.80 | 0.56 |
| | High | 2163 | 7.08 | 0.88 | 6.85 | 0.81 |
| 20 | Low | 5412 | 9.20 | 1.29 | 9.49 | 1.34 |
| | Low-Mod | 1299 | 10.72 | 0.83 | 10.93 | 0.74 |
| | Mod-High | 1176 | 9.91 | 0.78 | 9.78 | 0.65 |
| | High | 1788 | 7.09 | 0.90 | 6.93 | 0.87 |
| 25 | Low | 7264 | 9.67 | 1.49 | 10.06 | 1.61 |
| | Low-Mod | 724 | 11.65 | 0.96 | 11.90 | 0.85 |
| | Mod-High | 792 | 10.61 | 0.86 | 10.61 | 0.70 |
| | High | 1758 | 7.22 | 1.03 | 7.09 | 0.97 |
| 30 | Low | 8495 | 10.13 | 1.68 | 10.56 | 1.82 |
| | Low-Mod | 621 | 12.32 | 0.99 | 12.68 | 0.84 |
| | Mod-High | 656 | 11.18 | 1.01 | 11.31 | 0.77 |
| | High | 1464 | 7.14 | 0.95 | 7.04 | 0.94 |
| 35 | Low | 7041 | 10.60 | 1.94 | 11.06 | 2.11 |
| | Low-Mod | 455 | 12.90 | 0.93 | 13.47 | 0.90 |
| | Mod-High | 547 | 11.89 | 1.02 | 12.00 | 0.85 |
| | High | 1477 | 7.16 | 0.93 | 7.11 | 0.94 |
| 40 | Low | 4 | 14.09 | 0.83 | 15.99 | 0.16 |
| | Low-Mod | 282 | 13.47 | 1.00 | 13.94 | 1.03 |
| | Mod-High | 332 | 12.31 | 0.94 | 12.33 | 0.75 |
| | High | 1393 | 7.14 | 0.89 | 7.07 | 0.93 |

Table 5.40: Parameter estimates from Cox proportional hazards model for day final cumulative number infected for graphs above the epidemic threshold

| **Closeness Centralization** | | | | |
|---|---|---|---|---|
| **Day Final Cumulative Number Infected Graphs Above Epidemic Threshold** | | | | |
| **Parameter** | **Coefficient** | **SE** | **HR** | **p-value** |
| Centralization | 1.2190 | 0.0107 | 3.3837 | $< 0.0001$ |
| Most Central Node First (ref: Random) | -0.2448 | 0.0038 | 0.7829 | $< 0.0001$ |
| Order of Largest Component | -0.0943 | 0.0002 | 0.9100 | $< 0.0001$ |
| Centralization*Most Central Node First | 0.6204 | 0.0146 | 1.8597 | $< 0.0001$ |

## Other Analytic Approaches

| Closeness Centralization, Random Node First | | | | | | |
|---|---|---|---|---|---|---|
| Graph Order | | | GEE Model 2: Aggregate by graph | | GEE Model 1: No aggregation | |
| | Parameter | | Estimate | SE | Estimate | SE |
| 5 | Intercept | | 0.649 | 0.003 | 0.243 | 0.005 |
| | Closeness Centralization Quartile | Low-Mod | -0.085 | 0.003 | 0.145 | 0.005 |
| | | Mod-High | -0.113 | 0.003 | 0.192 | 0.005 |
| | | High | -0.110 | 0.003 | 0.188 | 0.004 |
| | Time | | 0.013 | 0.000 | 0.003 | 0.000 |

Figure 5.6: Parameter estimates from two GEE models predicting total number of nodes infected in the full network by closeness centralization quartile for 5 node graph.

| GEE Model Parameter Estimates [1,2] | | | | |
|---|---|---|---|---|
| *Random Node First* | | | | |
| | | Graph Order | | |
| Parameter | | 5 | 10 | 15 |
| Intercept | | 0.602 | 0.442 | 0.355 |
| Degree Centralization Quartile | Low-Mod | -0.019 | -0.085 | -0.102 |
| | Mod-High | -0.023 | -0.105 | -0.120 |
| | High | -0.057 | -0.127 | -0.146 |
| Time | | 0.013 | 0.017 | 0.017 |
| *Most Central Node First* | | | | |
| Intercept | | 0.572 | 0.382 | 0.287 |
| Degree Centralization Quartile | Low-Mod | -0.025 | -0.081 | -0.092 |
| | Mod-High | -0.028 | -0.102 | -0.114 |
| | High | -0.053 | -0.118 | -0.133 |
| Time | | 0.015 | 0.021 | 0.022 |

1 Total number of infected nodes predicted by degree centralization quartile (ref: low) and time

2 p<0.0001 for all estimates

Figure 5.7: Parameter estimates from GEE model predicting total number of nodes infected in the full network by degree centralization quartile for graphs of order 5-15

## Other Infection Parameters

Movie 5.29. Daily average total number of infected nodes in the largest component by degree centralization quartile with random node infected first for 30 node graph, probability of recovery 0.2, and probability of transmission between 0.7 and 0.9.

(Loading Video...)

Movie 5.30. Daily average total number of infected nodes in the largest component by degree centralization quartile with most central node infected first for 30 node graph, probability of recovery 0.2, and probability of transmission between 0.7 and 0.9.

(Loading Video...)

Movie 5.31. Daily average total number of infected nodes in the largest component by degree centralization quartile and type of node infected first for 30 node graph, probability of recovery 0.2, and probability of transmission between 0.7 and 0.9.

(Loading Video...)

Movie 5.32. Daily average cumulative number of infected nodes in the largest component by degree centralization quartile with random node infected first for 30 node graph, probability of recovery 0.2, and probability of transmission between 0.7 and 0.9.

(Loading Video...)

Movie 5.33. Daily average cumulative number of infected nodes in the largest component by degree centralization quartile with most central node infected first for 30 node graph, probability of recovery 0.2, and probability of transmission between 0.7 and 0.9.

(Loading Video...)

Movie 5.34. Daily average cumulative number of infected nodes in the largest component by degree centralization quartile and type of node infected first for 30 node graph, probability of recovery 0.2, and probability of transmission between 0.7 and 0.9.

(Loading Video...)

# Part V

# Conclusion

# Chapter 6

# Conclusion

Network science can aid in understanding the structure of the world around us. Network science concepts are particularly useful in providing insight into public health issues. For example, node centrality is commonly reported for public health networks to describe the relative importance of individuals based on their location in the network. Frequently used measures of centrality are closeness, betweenness, and degree. Closeness can be considered as a measure of node independence or efficiency of information transfer. Betweenness can be considered as a measure of node control. Lastly, degree is a measure of the number of direct contacts of a node.

The centrality measures of closeness, betweenness, and degree can be extended to centralization. The concept of centralization, or how much one node could dominate a network, seems intuitively important in public health networks. One node dominates all of the other nodes while in decentralized networks all nodes are equivalent. Highly centralized networks could be ideal for more rapid transmission of infections than decentralized networks. Unfortunately, very little research into centralization has been conducted. As a result, centralization is not commonly used to describe networks in the scientific literature. This program of research aimed to explore the properties of centralization and how it might be important in public health networks.

An examination of common graph-generating mechanisms found that none of the methods

currently used, including Erdös-Rènyi $Gnm$ random, Barabási-Albert preferential attachment, small world, and two node preferential attachment produce graphs along the full range of centralization values and with a variety graph structures. This limitation of established graph-generating methods stimulated the development of the Star Start program. As shown in Chapter 3, the Star Start method successfully produces graphs along the full range of centralization values and a wide variety of graph structures. As a result, the Star Start method is the ideal method to generate graphs for examination of centralization properties and to evaluate whether centralization could be important in disease spread networks. An examination of centralization properties was described in Chapter 4. Unfortunately, centralization is not clearly associated with summary centrality measures, such as maximum centrality or average centrality, that are easily calculated by common network software. Additionally, there is not a strong correlation between the three centralization measures examined. Chapter 5 examined $SIR$ models in networks produced using a modified version of the Star Start program that made it slightly more likely to produce centralized graphs. This research clearly shows that centralization is associated with key epidemiologic endpoints like peak number of infected nodes, cumulative number of infected nodes, and day that the peak number of nodes are infected.

Results of this research suggest that for the graph generating methods considered, low to moderate centralization levels are most commonly produced. Although highly centralized graphs are very unlikely to occur randomly, they have a profound impact on the course of an epidemic. Unfortunately, the distribution of centralization values for real networks is unknown since centralization is not commonly reported in the scientific literature. Furthermore, the forces that shape real networks are not well understood and there may be factors encouraging real networks to be more centralized than would be expected. Thus, a graph generating method that reliably produces highly centralized graphs, such as Star Start, can be used to investigate the influence of centralization on public health networks.

# Part VI

# Appendix

# Chapter 7

# A New Method for Creating Centralized Graphs: Star Start

## 7.1 Introduction

There are many different graph generating methods in the literature. The majority of these different methods aim to produce graphs with certain properties. For example, the small world method produces graphs with a high clustering coefficient and a short path length.[90] Barabási-Albert preferential attachment method produces graphs with a power law distribution for node degree.[87, 88] Erdös-Rènyi *Gnm* method randomly places edges in an empty graph and as a result produces a range of graph structures that are all equally likely.[86] However, if the interest is in producing graphs with the full range of centralization values and a range of graph structures these methods are inadequate. It has previously been shown that the small world method does not produce highly centralized graphs, Barabási-Albert preferential attachment method produces highly centralized graphs but with a limited range of structures.[97] Erdös-Rènyi *Gnm* random graphs do produce graphs with the full range of centralization values but doing so requires a very, very large number of graphs for even small graph orders.[149]

The Star Start program was designed to overcome these limitations by randomly sampling

from the range of possible Erdös-Rènyi $Gnm$ random graphs. In this way, Star Start provides an alternative method for generating graphs with the full range of centralization values that is efficient. Furthermore, the Star Start program can be used to approximate Erdös-Rènyi $Gnm$ random graphs so that inferences can be drawn.[149] Lastly, the Star Start program can be modified so that large numbers of moderate to highly centralized graphs are produced by limiting the number of graph changes that are allowed.

## 7.2 Star Start Program

The Star Start program begins by creating a complete graph ($g_0$) with the user-specified number of nodes, $n$. The edgelist of the complete graph is computed. Note, that this edgelist is the list of all possible edges that could be added to an empty graph with $n$ nodes. The list contains $\frac{n(n-1)}{2}$ unique edges. Next, a star graph ($g_1$) with the $n$ nodes is created. The edgelist of the star graph is computed (the Delete List) and compared to the edgelist of the complete graph so that a list of edges not in the star graph is produced (the Add List). Additionally, measurements of any graph and node properties of interest are taken (Star Start was originally designed to compute relative closeness, betweenness, and degree centrality and centralization[6], but other measures could be computed). A random number between 0 and 1 is generated. If the random number is $> 0.5$, an edge from the Delete List is randomly chosen and deleted from the graph. If the random number is $\leq 0.5$, an edge from the Add List is randomly selected and added to the graph. In this way, the structure of the next graph, $g_2$, is completely random and also prevents loops or multiple edges from being created. Again, graph and node property measurements are taken of the new graph as well a new list of edges that are not in the current graph. The process of randomly adding or deleting edges is continued until the graph is complete (contains all possible edges) or empty (contains no edges). As described here, there is no restriction on connectedness although the method can be adapted to meet such a restriction. The program reruns the Star Start method for a specified number of iterations.

The program outputs several different comma delimited files for the user to subsequently

analyze. Summary information as well as the actual centrality and centralization values for all graphs produced is output. The summary information includes minimum, mean, maximum centrality and number of graphs that fell into a particular centralization bin. The converse file with centrality binned is also produced. Another file records all of the unique maximum centrality values that fall into a particular centralization bin. The actual centrality and centralization value files distinguish different iterations of the programs in separate columns.

The program was developed in R version 2.15.1 [73] using the *igraph* version 0.6.2package [72].

## 7.3 Methods

As discussed in *Chapter 3: Centralization in Various Graph Generating Methods*, Star Start and Two Node Preferential Attachment both generate the full range of centralization values and graph structures.[97] In order to compare the utility of the Star Start Program with Two Node Preferential Attachment method in producing graphs with high centralization values, graphs were produced using both methods and the Freeman relative centrality and centralization measures for closeness, betweenness, and degree were computed for each graph.[6]

**Closeness** Intuitively, closeness between two nodes is simply a count of how many edges connect one particular node to another node in the graph. Let $d(n_i, n_j)$ be the number of edges in the geodesic between node $n_i$ and node $n_j$. By convention, if node $n_i$ and node $n_j$ are not connected by any edges, then $d(n_i, n_j) = \infty$. Of course, the number of edges between node $n_i$ and itself is 0, so $d(n_i, n_i) = 0$. Then the relative closeness centrality (called *closeness centrality* from here forward) of node $n_i$ (in a connected graph) is defined as $C'_c(n_i) = \left[\frac{\sum_{j=1}^{n} d(n_i,n_j)}{n-1}\right]^{-1} = \frac{n-1}{\sum_{j=1}^{n} d(n_i,n_j)}$. By definition, closeness centrality can only be computed for a connected graph (where each node is connected by at least one edge). However, *igraph* package in R substitutes $n$ for $\infty$ in the case of disconnected nodes so all

nodes have closeness centrality values. Consequently, for this paper closeness centrality is computed for all graphs, connected or disconnected.

Relative closeness centralization (called *closeness centralization* from here forward) is: $C_c = \frac{\sum_{i=1}^{n}[C'_c(n^*)-C'_c(n_i)]}{(n^2-3n+2)/(2n-3)}$, where $C'_c(n^*)$ is the maximum closeness centrality of the observed graph. Given this formula, it is clear that a centralization of 0 is obtained by a graph where all nodes are equal (ex: ring graph, empty graph, complete graph).

**Betweenness** Any nodes on the geodesic connecting two nodes are said to be between them. Betweenness centrality measures how often a node is between other nodes. $C_B(n_k) = \sum_{i}^{n}\sum_{<j}^{n} b_{ij}(n_k)$ where $b_{ij}(n_k) = \frac{g_{ij}(n_k)}{g_{ij}}$ is the number of geodesics connecting $n_i$ and $n_j$ that contain $n_k$ and $g_{ij}$ is the total number of geodesics connecting $n_i$ and $n_j$. Since the center point in a star graph obtains the maximum value, relative betweenness centrality (*betweenness centrality*) is defined as: $C'_B(n_k) = \frac{2*C_B(n_k)}{n^2-3n+2}$.

Relative betweenness centralization (*betweenness centralization*) is calculated as: $C_B = \frac{\sum_{i=1}^{n}[C'_B(n^*)-C'_B(n_i)]}{n-1}$ where $C'_B(n^*)$ is the maximum betweenness centrality of the observed graph.

**Degree** The number of edges connected to a node is the degree of the node. Degree centrality is defined as: $C_D(n_k) = \sum_{i=1}^{n} a(n_i, n_k)$ and relative degree centrality (*degree centrality*) is defined as: $C'_D(n_k) = \frac{\sum_{i=1}^{n} a(n_i,n_k)}{n-1}$ where

$$a(n_i, n_k) = \begin{cases} 1, \text{ iff edge between } n_i \text{ and } n_k \\ 0, \text{ otherwise} \end{cases}.$$

Relative degree centralization (*degree centralization*) is calculated as:
$C_D = \frac{\sum_{i=1}^{n}[C_D(n^*)-C_D(n_i)]}{n^2-3n+2}$ where $C_D(n^*)$ is the maximum degree centrality of the observed graph.

**Two Node Type Preferential Attachment** Heterogeneity in preference for forming edges between nodes of different types is used to understand the importance of node communities in a larger network and is the basis for homophily, or the idea that similar nodes

will link to each other.[91] In this method, nodes are randomly divided into a set number of types that link to each other with different probabilities. This paper investigated an equal probability of being either of two node types. The probability of a link forming between two nodes of the same type was $p_1$ and the probability of two nodes of different types forming a link was $p_2$. Both $p_1$ and $p_2$ ranged from 0 to 1 in increments of 0.01 and all combinations of $p_1$ and $p_2$ were tried. For each combination of values, 500 graphs were produced and the three measures of centrality and centralization calculated. Thus, 10,201 probability combinations were computed with 500 graphs produced for each combination. Note that two node type preferential attachment graphs are not required to be connected. Unlike the static version used here, most versions of node type preferential attachment add nodes and edges over time.[92, 93, 91]

**Star Start** The Star Start method was implemented as described above for 500 iterations. All three measures of centrality and centralization were taken for each update of $g$. Note that the number of graphs in each iteration is not fixed and as a result the a fixed number of iterations of the Star Start program produces a variable number of graphs.

## 7.4 Results

The Star Start program runs reasonably well for graphs of order 5 through 20 and 500 iterations. The program easily runs on a personal computer for graphs up to 10 nodes but thereafter memory problems occur due to the large number of graphs produced. Unfortunately, CPU time was not captured by the program for each graph order although the actual length of time for the program to run on a high performance computer cluster was recorded. Figure 7.1 describes the length in minutes of the actual computing time.

Figure 7.2 describes the total number of graphs produced by the program for 500 iterations at each graph order as well as the minimum, mean, and maximum number of graphs produced in a single iteration. As shown in the table, the number of graphs produced increases as graph order increases. Figure 7.3 describes the number of graphs that exited an iteration

as either complete (fully connected) or empty (fully disconnected) by graph order.

As discussed in *Chapter 3: Centralization in Various Graph Generating Methods*, Star Start and Two Node Type Preferential Attachment both generate the full range of centralization values and graph structures.[97] As shown in Movies 7.1-7.3, the distribution of graphs into four centralization categories (called low, low-moderate, moderate-high, and high) based on centralization quartile are similar for both Two Node Type Preferential Attachment and Start Start. Further comparison of the these two programs suggests that the Star Start program can more efficiently produce many graphs of moderate to high centralization values. As shown in Figures 7.4-.6, a relatively small number of graph changes captures the majority of moderate to high centralization values. The results suggest that for all three centralization measures, the maximum number of graph changes can be set to be $2 * n$, where $n$ is number of nodes, in order to capture more than 50% of graphs with centralization $\geq 0.3$. Figures 7.7-.9 illustrate the top five combinations of node attachment probabilities for the Two Node Type Preferential Attachment model that generate the highest percentage of highly centralized graphs for graphs of order 5-8. As shown in the tables, certain combinations of node attachment probabilities are more likely to produce highly centralized graphs across graph order. However, none of these combinations produce highly centralized graphs more than 40% of the time and the probability of producing a highly centralized graph decreases as graph order increases.

## 7.5 Discussion

As described previously, the Star Start program produces most of the full range of centralization values for closeness, betweenness, and degree centralization (see [97]). Also, in order to complete a Star Start iteration, either a minimum of $n-1$ edges must be removed from the star graph to create an empty graph or a minimum of $\frac{(n-1)(n-2)}{2}$ edges must be added to the star graph to create a complete graph.[97] The minimum number of graphs information in Table 1 confirms that the program is working according to design. The fact that fewer steps are required to exit the iteration through creating an empty graph is clearly

illustrated by the increasing number of empty graphs as graph order increases.

Although the number of graphs produced for Star Start is quite large, it is much less than the $2^{\frac{n(n-2)}{2}}$ graphs that would be required to produce the distribution of centralization according to Erdös-Rènyi $Gnm$ random graphs for a set number of nodes (see [149] for a comparison of the graph numbers required for graphs of order 5-8).

A comparison of the utility of the Star Start program and Two Node Type Preferential Attachment in producing highly centralized graphs suggests that with some modification, the Star Start program can easily produce large numbers of highly centralized graphs. Unlike the Two Node Type Preferential Attachment method which produces graphs based on node linkage probabilities, the Star Start program starts with the most centralized graph (a star graph) and with each step produces a new graph by adding or deleting an edge. In this way, the number of graph changes for each Star Start iteration can be limited to obtain large numbers of highly centralized graphs without also producing large numbers of low centralization graphs. Unfortunately, even when the combinations of node attachment probabilities that are most likely to produce highly centralized graphs are used, the vast majority of those graphs have low centralization.

## 7.6 Conclusion

The Star Start program provides a new method to produce graphs that span most of the full range centralization values for graphs of order 5-20. Additionally, the Star Start program provides a reasonable alternative to computing a large number of Erdös-Rènyi $Gnm$ random graphs for inference purposes. Lastly, the Star Start program provides an efficient way to generate large numbers of highly centralized graphs.

## 7.7 Figures, Movies

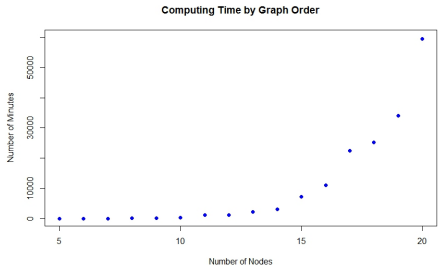


Figure 7.1: Figure 1. Actual computing time for Star Start program by graph order

| Order of graph (# of nodes) | Total # graphs generated | Min # graphs in an iteration | Mean # graphs in an iteration | Max # graphs in an iteration |
|---|---|---|---|---|
| 5 | 12474 | 5 | 25 | 109 |
| 6 | 24960 | 6 | 50 | 282 |
| 7 | 43857 | 7 | 88 | 453 |
| 8 | 75126 | 8 | 150 | 976 |
| 9 | 118558 | 9 | 237 | 1511 |
| 10 | 151563 | 10 | 303 | 3217 |
| 11 | 219058 | 13 | 438 | 3714 |
| 12 | 304192 | 14 | 608 | 4496 |
| 13 | 402246 | 23 | 804 | 9365 |
| 14 | 426436 | 18 | 853 | 6642 |
| 15 | 693192 | 21 | 1386 | 10742 |
| 16 | 814884 | 24 | 1630 | 13832 |
| 17 | 1020840 | 23 | 2042 | 22869 |
| 18 | 1178231 | 28 | 2356 | 23377 |
| 19 | 1310591 | 37 | 2621 | 27772 |
| 20 | 1772296 | 54 | 3545 | 39412 |

Figure 7.2: Number of graphs produced by the Star Start program by graph order

| Order of graph (# of nodes) | # Empty graphs | # Complete graphs |
|---|---|---|
| 5 | 312 | 188 |
| 6 | 332 | 168 |
| 7 | 353 | 147 |
| 8 | 367 | 133 |
| 9 | 380 | 120 |
| 10 | 409 | 91 |
| 11 | 408 | 92 |
| 12 | 428 | 72 |
| 13 | 430 | 70 |
| 14 | 440 | 60 |
| 15 | 412 | 88 |
| 16 | 432 | 68 |
| 17 | 436 | 64 |
| 18 | 449 | 51 |
| 19 | 445 | 55 |
| 20 | 454 | 46 |

Figure 7.3: Number of graphs with either empty or complete structure at the end of a Star Start iteration by graph order

Movie 7.1. Closeness centralization quartile for Star Start and Two Node Preferential Attachment obtained for graphs of order 5-20 nodes.

(Loading Video...)

Movie 7.2. Betweenness centralization quartile for Star Start and Two Node Preferential Attachment obtained for graphs of order 5-20 nodes.

(Loading Video...)

Movie 7.3. Degree centralization quartile for Star Start and Two Node Preferential Attachment obtained for graphs of order 5-20 nodes.

(Loading Video...)

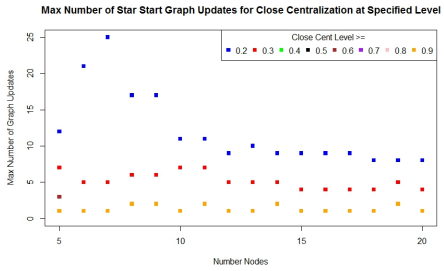


Figure 7.4: Maximum number of graph changes required to produce at least 50% of graphs with closeness centralizations $\geq$ a specified level

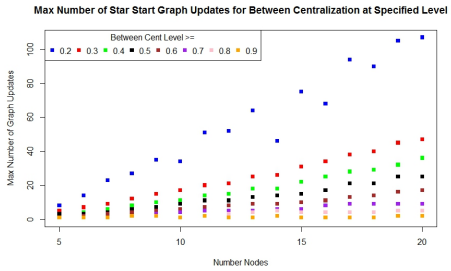


Figure 7.5: Maximum number of graph changes required to produce at least 50% of graphs with betweenness centralizations $\geq$ a specified level

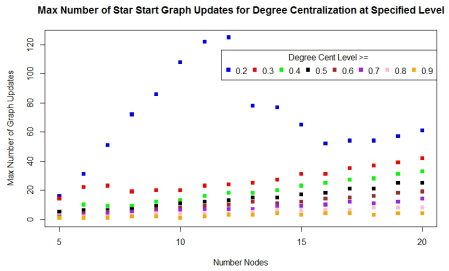


Figure 7.6: Maximum number of graph changes required to produce at least 50% of graphs with degree centralizations $\geq$ a specified level

| Graph Order (# nodes) | Node Attachment Probability | | Closeness Centralization Category Relative Frequency | | | |
|---|---|---|---|---|---|---|
| | Same Type | Diff Type | Low | Low-Mod | Mod-High | High |
| 5 | 0.00 | 1.00 | 0.64 | NA | NA | 0.36 |
| | 0.04 | 0.97 | 0.52 | 0.11 | 0.02 | 0.35 |
| | 0.14 | 0.98 | 0.50 | 0.10 | 0.06 | 0.34 |
| | 0.08 | 0.97 | 0.51 | 0.11 | 0.04 | 0.34 |
| | 0.11 | 0.99 | 0.54 | 0.07 | 0.06 | 0.33 |
| 6 | 0.17 | 0.99 | 0.38 | 0.27 | 0.13 | 0.22 |
| | 0.06 | 1.00 | 0.35 | 0.39 | 0.04 | 0.22 |
| | 0.00 | 1.00 | 0.33 | 0.46 | NA | 0.21 |
| | 0.03 | 1.00 | 0.33 | 0.44 | 0.02 | 0.21 |
| | 0.10 | 1.00 | 0.34 | 0.39 | 0.06 | 0.20 |
| 7 | 0.00 | 1.00 | 0.53 | 0.34 | NA | 0.13 |
| | 0.03 | 0.98 | 0.48 | 0.32 | 0.06 | 0.13 |
| | 0.10 | 0.99 | 0.39 | 0.42 | 0.06 | 0.13 |
| | 0.15 | 1.00 | 0.34 | 0.48 | 0.06 | 0.12 |
| | 0.07 | 1.00 | 0.45 | 0.40 | 0.03 | 0.12 |
| 8 | 0.10 | 0.98 | 0.45 | 0.31 | 0.14 | 0.09 |
| | 0.23 | 0.99 | 0.35 | 0.46 | 0.10 | 0.09 |
| | 0.04 | 1.00 | 0.67 | 0.06 | 0.18 | 0.09 |
| | 0.11 | 0.99 | 0.49 | 0.29 | 0.14 | 0.09 |
| | 0.07 | 0.96 | 0.49 | 0.27 | 0.16 | 0.08 |

Figure 7.7: Top five combinations of node attachment probabilities for the Two Node Type Preferential Attachment model that produce the highest percentage of graphs with high closeness centralization for graphs of order 5-8 nodes

| Graph Order (# nodes) | Node Attachment Probability | | Betweenness Centralization Category Relative Frequency | | | |
|---|---|---|---|---|---|---|
| | Same Type | Diff Type | Low | Low-Mod | Mod-High | High |
| 5 | 0.00 | 1.00 | 0.64 | NA | NA | 0.36 |
| | 0.04 | 0.97 | 0.54 | 0.11 | 0.01 | 0.34 |
| | 0.01 | 1.00 | 0.67 | NA | NA | 0.33 |
| | 0.00 | 0.99 | 0.64 | 0.03 | NA | 0.33 |
| | 0.04 | 0.98 | 0.59 | 0.07 | 0.01 | 0.32 |
| 6 | 0.00 | 1.00 | 0.79 | NA | NA | 0.21 |
| | 0.06 | 1.00 | 0.78 | NA | 0.01 | 0.21 |
| | 0.03 | 1.00 | 0.79 | NA | 0.00 | 0.20 |
| | 0.05 | 1.00 | 0.81 | NA | 0.00 | 0.19 |
| | 0.07 | 1.00 | 0.80 | 0.00 | 0.02 | 0.18 |
| 7 | 0.00 | 1.00 | 0.53 | 0.34 | NA | 0.13 |
| | 0.03 | 0.98 | 0.60 | 0.26 | 0.01 | 0.13 |
| | 0.03 | 1.00 | 0.63 | 0.25 | NA | 0.12 |
| | 0.10 | 0.99 | 0.72 | 0.14 | 0.01 | 0.12 |
| | 0.05 | 1.00 | 0.68 | 0.20 | 0.00 | 0.12 |
| 8 | 0.04 | 1.00 | 0.73 | 0.19 | NA | 0.08 |
| | 0.10 | 0.98 | 0.74 | 0.16 | 0.02 | 0.08 |
| | 0.02 | 1.00 | 0.70 | 0.23 | NA | 0.07 |
| | 0.03 | 1.00 | 0.75 | 0.18 | NA | 0.07 |
| | 0.10 | 1.00 | 0.80 | 0.13 | NA | 0.07 |

Figure 7.8: Top five combinations of node attachment probabilities for the Two Node Type Preferential Attachment model that produce the highest percentage of graphs with high betweenness centralization for graphs of order 5-8 nodes

| Graph Order (# nodes) | Node Attachment Probability | | Degree Centralization Category Relative Frequency | | | |
|---|---|---|---|---|---|---|
| | Same Type | Diff Type | Low | Low-Mod | Mod-High | High |
| 5 | 0.00 | 1.00 | 0.64 | NA | NA | 0.36 |
| | 0.04 | 0.97 | 0.51 | 0.11 | 0.03 | 0.34 |
| | 0.01 | 1.00 | 0.66 | 0.01 | NA | 0.33 |
| | 0.00 | 0.99 | 0.64 | 0.02 | 0.01 | 0.33 |
| | 0.04 | 0.98 | 0.57 | 0.08 | 0.04 | 0.32 |
| 6 | 0.06 | 1.00 | 0.34 | 0.41 | 0.03 | 0.22 |
| | 0.00 | 1.00 | 0.33 | 0.46 | NA | 0.21 |
| | 0.03 | 1.00 | 0.33 | 0.44 | 0.02 | 0.21 |
| | 0.00 | 0.85 | 0.27 | 0.46 | 0.07 | 0.20 |
| | 0.01 | 0.91 | 0.34 | 0.43 | 0.03 | 0.20 |
| 7 | 0.03 | 0.98 | 0.49 | 0.32 | 0.06 | 0.14 |
| | 0.00 | 1.00 | 0.53 | 0.34 | NA | 0.13 |
| | 0.02 | 0.96 | 0.49 | 0.29 | 0.08 | 0.13 |
| | 0.10 | 0.99 | 0.45 | 0.37 | 0.05 | 0.13 |
| | 0.02 | 0.94 | 0.44 | 0.31 | 0.13 | 0.13 |
| 8 | 0.07 | 0.96 | 0.53 | 0.22 | 0.16 | 0.09 |
| | 0.10 | 0.98 | 0.49 | 0.28 | 0.14 | 0.09 |
| | 0.04 | 1.00 | 0.68 | 0.06 | 0.18 | 0.08 |
| | 0.11 | 0.99 | 0.53 | 0.26 | 0.13 | 0.08 |
| | 0.04 | 0.95 | 0.40 | 0.30 | 0.22 | 0.08 |

Figure 7.9: Top five combinations of node attachment probabilities for the Two Node Type Preferential Attachment model that produce the highest percentage of graphs with high degree centralization for graphs of order 5-8 nodes

## 7.8 Program

This is the main Star Start program.

```
AddMinStar4=function(ver,rep)
  # ver is number of vertices for the initial star graph
  # rep is number of iterations
  # type is "closeness", "betweeess", "degree"
{
  library(igraph)

  if(!exists("matrixFromList.R")){
    source("C:/Users/Christina/NetworkResearch/matrixFromList.R")}
  if(!exists("gprop3.R")){
    source("C:/Users/Christina/NetworkResearch/gprop3.R")}

  # need lists here since do not know how long vectors will be
  CClose=list() # capture maximum vertex closeness centrality information
  CVClose=list() # capture all vertex closeness centralization information
  CCentral=list() # capture network closeness centralization information
```

```
  BClose=list() # capture maximum vertex betweenness centrality information
  BVClose=list() # capture all vertex betweenness centrality information
  BCentral=list() # capture network betweenness centralization information

  DClose=list() # capture maximum vertex degree centrality information
  DVClose=list() # capture all vertex degree centrality information
  DCentral=list() # capture network degree centralization information

  ranlist=list() # want a list to capture all random numbers for each iteration

  for (j in 1:rep){  # do this for every iteration

    g=list() # need a list to capture the networks
    ranlist[[j]]=list() # need a list to capture random numbers
    g[[1]]=graph.star(ver,mode="undirected") # starting network is star
    e=get.edgelist(graph.full(ver, directed=FALSE, loops=FALSE))
     # need edgelist for complete graph for comparison
    b=get.edgelist(g[[1]]) # edgelist of first graph set
    edge=e[apply(e, 1, function(x)
max( apply(b, 1, function(y)all.equal(x, y,
     check.attributes=FALSE)))) != "TRUE",]
    i=1
    ranlist[[j]][[1]]=NULL
    # loop while edgelist has edges to add
    while (length(edge) >0 & length(edge)<length(e)){

      i=i+1 # update graph number

      ranlist[[j]][[i]]=runif(1) # generate random number to decide to add
```

```
       #or remove edges
      if (ranlist[[j]][[i]]>0.5){ # remove edges
        g[[i]]=delete.edges(g[[(i-1)]],
E(g[[i-1]],c(b[c(sample(1:nrow(b),1)),])))
         # remove the randomly selected edge
      }else if(ranlist[[j]][[i]]<=0.5 & length(edge)!=2){
        g[[i]]=add.edges(g[[(i-1)]],edge[sample(1:nrow(edge),1),])
         # randomly add edges by choosing two vertices
      }else if(ranlist[[j]][[i]]<=0.5 & length(edge)==2){ # add edges
        g[[i]]=add.edges(g[[i-1]],edge)}
        # randomly add edges by choosing two vertices

      b=get.edgelist(g[[i]],names=TRUE)
# need current edgelist to know what
       #can be added/removed
      # get list of edges that can be added

      if (length(b)>2){
        edge=e[apply(e, 1, function(x) max( apply(b, 1,
         function(y)all.equal(x, y,
         check.attributes=FALSE)))) != "TRUE",]
      }else if (length(b)==2){
        edge=e[as.vector(1-apply(e,1,function(x)
        all(x==b,arr.ind=TRUE)),mode="logical"),]
      }else{edge=e}

    } # end loop for adding/deleting vertices

    # CLOSENESS centrality/centralization for the networks produced
#for this iteration
```

```
    Close=sapply(g,function(g)
     {centralization.closeness(g,mode="all",normalized=TRUE)})
     # all closeness values
    CCentral[[j]]=as.numeric(Close[2,]) # closeness centralization for each graph
    CVClose[[j]]=Close[1,] # vertex closeness centrality
    CClose[[j]]=as.numeric(lapply(CVClose[[j]],max))
# maximum vertex closeness centrality

    # BETWEENNESS centrality/centralization for the networks produced
#for this iteration
    Betw=sapply(g,function(g)
     {centralization.betweenness(g,normalized=TRUE)})
     # all betweenness values
    BCentral[[j]]=as.numeric(Betw[2,]) # betweenness centralization for each graph
    BVClose[[j]]=t(sapply(g,function(g){betweenness(g,normalized=TRUE)}))
    BClose[[j]]=apply(BVClose[[j]],1,max) # maximum vertex betweenness centrality

    # DEGREE centrality/centralization for the networks produced
#for this iteration
    Deg=sapply(g,function(g)
     {centralization.degree(g,mode="total",loops=FALSE,normalized=TRUE)})
     # all degree values
    DCentral[[j]]=as.numeric(Deg[2,]) # degree centralization for each graph
    DVClose[[j]]=t(sapply(g,function(g)
     {degree(g,mode="total",loops=FALSE,normalized=TRUE)}))

    DClose[[j]]=apply(DVClose[[j]],1,max)

    # save graph information for each repetition in case processing gets interrupted
    write.table(t(CCentral[[j]]),
```

```
file=paste("C:/Users/Christina/NetworkResearch/ImpCation",
     ver,"close",rep,".txt",sep=""), append=TRUE,sep=",",
row.names=FALSE,col.names=FALSE)
    write.table(t(BCentral[[j]]),
file=paste("C:/Users/Christina/NetworkResearch/ImpCation",
     ver,"between",rep,".txt",sep=""),append=TRUE,sep=",",
row.names=FALSE,col.names=FALSE)
    write.table(t(DCentral[[j]]),
file=paste("C:/Users/Christina/NetworkResearch/ImpCation",
     ver,"degree",rep,".txt",sep=""), append=TRUE,sep=",",
row.names=FALSE,col.names=FALSE)
    write.table(CVClose[[j]],
file=paste("C:/Users/Christina/NetworkResearch/ImpVerCity",
     ver,"close",rep,".txt",sep=""), append=TRUE,sep=",",
row.names=FALSE,col.names=FALSE)
    write.table(t(BVClose[[j]]),
file=paste("C:/Users/Christina/NetworkResearch/ImpVerCity",
     ver,"between",rep,".txt",sep=""), append=TRUE,sep=",",
row.names=FALSE,col.names=FALSE)
    write.table(t(DVClose[[j]]),
file=paste("C:/Users/Christina/NetworkResearch/ImpVerCity",
     ver,"degree",rep,".txt",sep=""), append=TRUE,sep=",",
row.names=FALSE,col.names=FALSE)
  } # end repetitions loop

  gprop3(ver,rep,"closeness",CCentral,CClose)
  gprop3(ver,rep,"betweenness",BCentral,BClose)
  gprop3(ver,rep,"degree",DCentral,DClose)

} # end the function
```

```
# call the function
AddMinStar4(5,500)
```

This is the function that is called to produce the summary tables.

```
gprop3=function(ver,rep,type,AA,BB)
{
# AA is centralization matrix
# BB is centrality matrix

# now need to work with lists and summarize information
# for most things need matrices
mCentral=matrixFromList(AA) # centralization matrix
mClose=matrixFromList(BB) # centrality matrix

mCentral2=apply(mCentral,2,function(x){cut(round(x,digits=7),
breaks=seq(-0.01,1,by=0.01),right=TRUE)})
# categorize centralization matrix into bins
mClose2=apply(mClose,2,function(x){cut(round(x,digits=7),
breaks=seq(-0.01,1,by=0.01),right=TRUE)})
# categorize closeness centrality matrix into bins
  # need to round so that the "1" values bin properly.  Not sure why

cat=as.vector(cut(seq(-0.01,by=0.01),seq(-0.01,1,by=0.01))[-1])
# create 100 bins for centralization
indx= apply(mCentral2,MARGIN=2, FUN=function(x) {match(x, cat)})
# find the locations of each centralization bin
indx2= apply(mClose2,MARGIN=2, FUN=function(x) {match(x, cat)})
# find the locations of each centrality bin
```

```
# summarize the CENTRALITY information by CENTRALIZATION bin
lClose=list() # create list to store information
for (j in 1:101){ # loop for each centralization bin
  lClose[[j]]=mClose[which(indx==j,arr.ind=TRUE)]
  # list of all centrality values for a particular cenralization bin
} # end loop
mlClose=t(matrixFromList(lClose)) # create matrix from binned centrality list
graphClose=apply(mlClose,MARGIN=1,unique)
mgraphClose=t(matrixFromList(graphClose))
mgraphClose2=cbind(seq(0,1,by=0.01),mgraphClose)
# need centrality and centralization info to graph
rownames(mlClose)=cat # add row names so know bins

mlClose2=mlClose[!is.na(mlClose[,1]),] # get rid of bins with no information
mgraphClose3=mgraphClose2[!is.na(mgraphClose2[,2]),]
# get rid of bins with no information
colnames(mgraphClose3)=
c("centralization",sprintf("centrality%d",1:(ncol(mgraphClose2)-1)))

# summary information for centrality binned by centralization
CitybyCation=cbind(
apply(mlClose2, MARGIN=1,FUN=function(x) {min(as.numeric(x), na.rm=TRUE)}),
      apply(mlClose2, MARGIN=1,FUN=function(x) {mean(as.numeric(x), na.rm=TRUE)}),
      apply(mlClose2, MARGIN=1,FUN=function(x) {max(as.numeric(x), na.rm=TRUE)}),
      rowSums(!is.na(mlClose2)))

colnames(CitybyCation)=c("min", "mean" ,"max", "n") # column names

# summarize the CENTRALIZATION information by CENTRALITY bin
```

```
lCentral2=list() # create list to store information
for (j in 1:101){ # loop for each centralization bin
  lCentral2[[j]]=mCentral[which(indx2==j,arr.ind=TRUE)]
  #list of centralization values for a particular centralization bin
} # end loop
mlCentral3=t(matrixFromList(lCentral2))
# create matrix from binned centralization list
rownames(mlCentral3)=cat # add row names so know bins


mlCentral4=mlCentral3[!is.na(mlCentral3[,1]),]
# get rid of bins with no information


# summary information for centralization binned by centralization
CationbyCity=cbind(
apply(mlCentral4, MARGIN=1,FUN=function(x) {min(as.numeric(x), na.rm=TRUE)}),
    apply(mlCentral4, MARGIN=1,FUN=function(x) {mean(as.numeric(x), na.rm=TRUE)}),
    apply(mlCentral4, MARGIN=1,FUN=function(x) {max(as.numeric(x), na.rm=TRUE)}),
    rowSums(!is.na(mlCentral4)))

colnames(CationbyCity)=c("min", "mean" ,"max", "n") # column names
write.table(CitybyCation,paste("C:/Users/Christina/NetworkResearch/ImpAMCitybyCation",
ver,type, rep,".txt",sep=""), sep=",",row.names=TRUE,col.names=TRUE)
write.table(CationbyCity,paste("C:/Users/Christina/NetworkResearch/ImpAMCationbyCity",
ver,type,rep,".txt",sep=""), sep=",",row.names=TRUE,col.names=TRUE)
write.table(mgraphClose3,paste("C:/Users/Christina/NetworkResearch/ImpAMgCity",
ver,type,rep,".txt",sep=""), sep=",",row.names=FALSE,col.names=TRUE)
}
```

This is the function that creates a matrix out of a list.

```
# function to create matrix from list
matrixFromList=function(listX)
{
  sapply(listX, function(x, n) c(x, rep(NA, n))[1:n], n = max(sapply(listX, length)))}
```

# Chapter 8

# Bibliography

# Bibliography


[1] M.E.J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[2] Thomas W. Valente. *Social Networks and Health: Moels, Methods, and Applications*. Oxford University Press, 2010.

[3] Douglas A. Luke and Jenine K. Harris. "Network analysis in public health: History, methods, and applications". In: *Annual Review of Public Health* 28 (2007), pp. 69–93.

[4] Ted G. Lewis. *Network Science: Theory and Practice*. John Wiley and Sons, Inc., 2009.

[5] Ulrik Brandes and Daniel Fleischer. "Centrality measures based on current flow". In: *STACS: 22nd Annual Symposium on Theoretical Aspects of Computer Science*. Vol. 3404. Lecture Notes in Computer Science. Springer, 2005, pp. 533–544.

[6] Linton C. Freeman. "Centrality in Social Networks: Conceptual Clarification". In: *Social Networks* 1 (1978), 215:239.

[7] Tord Høivik and Nils Petter Gleditsch. "Structural parameters of graphs: A theoretical investigation". In: *Quality & Quantity* 4.1 (1970), pp. 193–209.

[8] M.G. Everett and P. Sinclair. "Some Centrality Results New and Old". In: *Journal of Mathematical Sociology* 28 (2004), 215:227.

[9] Stephen P. Borgatti and Martin G. Everett. "Network analysis of 2-mode data". In: *Social Networks* (1997), pp. 243–269.

[10] Dirk Koschützki et al. “Centrality indices”. In: *Network analysis*. Springer, 2005, pp. 16–61.

[11] Rodrigo A Botafogo, Ehud Rivlin, and Ben Shneiderman. “Structural analysis of hypertexts: identifying hierarchies and useful metrics”. In: *ACM Transactions on Information Systems (TOIS)* 10.2 (1992), pp. 142–180.

[12] T. W. Valente and K. Fujimoto. “Bridging: Locating critical connectors in a network”. In: *Social Networks* 32 (3 2010), pp. 212–220.

[13] Stephen P. Borgatti. “Centrality and network flow”. In: *Social Networks* 27 (2005), pp. 55–71.

[14] Murray A Beauchamp. “An improved index of centrality”. In: *Behavioral Science* 10.2 (1965), pp. 161–163.

[15] Claude Flament. *Applications of Graph Theory to Group Structure*. Prentice-Hall, Inc., 1963.

[16] Linton C. Freeman. “A set of measures of centrality based on betweenness”. In: *Sociometry* 40 (1 1977), pp. 35–41.

[17] Mark EJ Newman. “Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality”. In: *Physical Review E* 64.1 (2001), p. 016132.

[18] R. Grassi et al. “Betweenness centrality: Extremal values and structural properties”. In: *Lecture notes in economics and mathematical systems* 613 (2009), pp. 161–175.

[19] F. Comellas and S. Gago. “Spectral bounds for the betweenness of a graph”. In: *Linear Algebra and its Applications* 423 (2007), pp. 74–80.

[20] Juhani Nieminen. “On the centrality in a graph”. In: *Scandinavian Journal of Psychology* 15.1 (1974), pp. 332–336.

[21] Carter T. Butts. “Exact bounds for degree centralization”. In: *Social Networks* 28 (2006), pp. 283–296.

[22] Tim Dwyer et al. “Visual analysis of network centralities”. In: Australian Computer Society, Inc. 2006.

[23] Thomas W. Valente et al. “How correlated are network centrality measures?” In: *Connections: Bulletin of the International Network for Social Network Analysis* 28 (1 2008), pp. 16–26.

[24] Petter Holme et al. “Attack vulnerability of complex networks”. In: *Physical Review E* 65.5 (2002), p. 056109.

[25] Bonan Hou, Yiping Yao, and Dongsheng Liao. “Identifying all-around nodes for spreading dynamics in complex networks”. In: *Physica A: Statistical Mechanics and its Applications* 391 (2012), pp. 4012–4017.

[26] K-I Goh et al. “Betweenness centrality correlation in social networks”. In: *Physical Review E* 67.1 (2003), p. 017101.

[27] Keiko Nakao. “Distribution of measures of centrality: enumerated distributions of Freeman’s graph centrality measures”. In: *Connections: Bulletin of the International Network for Social Network Analysis* 13 (3 1990), pp. 10–22.

[28] Brigham S. Anderson, Carter Butts, and Kathleen Carley. “The interatction of size and density with graph-level indices”. In: *Social Networks* 21 (1999), pp. 239–267.

[29] Elizabeth Costenbader and Thomas W. Valente. “The stability of centrality measures when networks are sampled”. In: *Social Networks* 25 (2003), pp. 283–397.

[30] Stephen P. Borgatti, Kathleen M. Carley, and David Krackhardt. “On the robustness of centrality measures under conditions of imperfect data”. In: *Social Networks* 28 (2006), pp. 124–136.

[31] Søren Atmakuri Davidsen. *Sampling/misinformation effect on centrality measures in complex networks*. Tech. rep. Aalborg University Esbjerg, 2009.

[32] Barbara Zemljič and Valentina Hlebec. “Reliability of measures of centrality and prominence”. In: *Social Networks* 27.1 (2005), pp. 73–88.

[33] Duanbing Che et al. “Identifying influential nodes in complex networks”. In: *Physica A: Statistical Mechanics and its Applications* 391 (2012), pp. 1777–1787.

[34] N. E. Friedkin. “Theoretical foundations for centrality measures”. In: *American Journal of Sociology* (1991), pp. 1478–1504.

[35] Muhammad Akram Shaikh et al. “Graph Structural Mining in Terrorist Networks”. In: Springer-Verlag, 2007, pp. 570–577.

[36] Jorge Gil-Mendieta and Samuel Schmidt. “The political network in Mexico”. In: *Social Networks* 18.4 (1996), pp. 355–381.

[37] Ann-Marie Kermarrec et al. “Second order centrality: Distributed assessment of nodes criticity”. In: *Computer Communications* 34 (2011), pp. 619–628.

[38] Ernesto Estrada and Juan A. Rodríguez-Velázquez. “Subgraph centrality in complex networks”. In: *Physical Review E* 71 (2005), p. 056103.

[39] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. “Control centrality and hierarchical structure in complex networks”. In: *PloS One* 7.9 (2012), e44459.

[40] M.E.J. Newman. “A measure of betweenness centrality based on random walks”. In: *Social Networks* 27 (2005), pp. 39–54.

[41] Yoosik Youm. “A sociological interpretation of emerging properties in STI transmission dynamics: walk-betweenness of sexual networks”. In: *Sexually Transmitted Infections* 86.Suppl 3 (2010), pp. iii24–iii28.

[42] Karen Stephenson and Marvin Zelen. “Rethinking centrality: Methods and examples”. In: *Social Networks* (1989), pp. 1–37.

[43] Vito Latora and Massimo Marchiori. “Efficient behavior of small-world networks”. In: *Physical review letters* 87 (19 2001), pp. 198701–1–198701–4.

[44] Chavdar Dangalchev. “Residual closeness in networks”. In: *Physica A: Statistical Mechanics and its Applications* 365 (2006), pp. 556–564.

[45] Aysun Aytac and Zeynep Nihan Odabas. “Residual closeness of wheels and related networks”. In: *International Journal of Foundations of Computer Science* 22 (5 2011), pp. 1229–1240.

[46] D. J. Klein. “Centrality measure in graphs”. In: *Journal of Mathematical Chemistry* 47.4 (2010), pp. 1209–1223.

[47] Phillip Bonacich. “Factoring and weighting approaches to status scores and clique identification”. In: *Journal of Mathematical Sociology* 2.1 (1972), pp. 113–120.

[48] Phillip Bonacich. “Some unique properties of eigenvector centrality”. In: *Social Networks* 29.4 (2007), pp. 555–564.

[49] Britta Ruhnau. “Eigenvector centrality a node centrality?” In: *Social Networks* 22.4 (2000), pp. 357–365.

[50] Phillip Bonacich. “Simultaneous group and individual centralities”. In: *Social Networks* 13 (1991), pp. 155–168.

[51] Phillip Bonacich, Amalya Oliver, and Tom A.B. Snijders. “Controlling for size in centrality scores”. In: *Social Networks* 20 (1998), pp. 135–141.

[52] Phillip Bonacich. “Power and centrality: A family of measures”. In: *American Journal of Sociology* 92 (5 1987), pp. 1170–1182.

[53] Phillip Bonacich and Paulette Lloyd. “Eigenvector-like measures of centrality for asymmetric relations”. In: *Social Networks* 23 (2001), pp. 191–201.

[54] Christian Tallberg. “Testing centralization in random graphs”. In: *Social Networks* 26 (2004), pp. 205–219.

[55] S Gago, J Hurajova, and T Madaras. “Notes on the betweenness centrality of a graph”. In: *Mathematica Slovaca* 62 (1 2012), pp. 1–12.

[56] Philip Andrew Sinclair. “Network centralization with the Gil Schmidt power centrality index”. In: *Social Networks* 31 (2009), pp. 214–219.

[57] Vincent Buskens. “The social structure of trust”. In: *Social Networks* 20 (1998), pp. 265–289.

[58] Tom A.B. Snijders. “The degree variance: An index of graph heterogeneity”. In: *Social Networks* 3 (1981), pp. 163–174.

[59] Nicholas A. Christakis and James H. Fowler. “The collective dynamics of smoking in a large social network”. In: *The New England Journal of Medicine* 358 (21 2008), pp. 2249–2258.

[60] P De et al. “Sexual network analysis of a gonorrhoea outbreak”. In: *Sexually Transmitted Infections* 80 (2004), pp. 280–285.

[61] Peter S. Bearman, James Moody, and Katherine Stovel. "Chains of affection: The structure of adolescent romantic and sexual networks". In: *American Journal of Sociology* 110 (1 2004), pp. 44–91.

[62] Thomas Hornbeck et al. "Using sensor networks to study the effect of peripatetic healthcare workers on the spread of hospital-associated infections". In: *The Journal of Infectious Diseases* 206 (2012), pp. 1549–1557.

[63] Melissa Krauss, Nancy Mueller, and Douglas Luke. "Interorganizational relationships within state tobacco control networks: A social network analysis". In: *Preventing Chronic Disease: Public Health Research, Practice, and Policy* 1 (4 2004), pp. 1–25.

[64] Chung-Yuan Huang et al. "Simulating SARS: Small world epidemiological modeling and public health policy assessments". In: *Journal of Artificial Societies and Social Simulation* 7 (4 2004).

[65] Lauren Ancel Meyers et al. "Network theory and SARS: Predicting outbreak diversity". In: *Journal of Theoretical Biology* 232 (2005), pp. 71–81.

[66] E. Le Merrer and G. Trédan. "Centralities: capturing the fuzzy notion of importance in social graphs". In: 2009.

[67] Patrick Doreian and Kayo Fujimoto. "Identifying linking-pin organizations in interorganization networks". In: *Computational and Mathematical Organization Theory* 10 (2004), pp. 45–68.

[68] Curtis M. Topper and Kathleen M. Carley. "A structural perspective on the emergence of network organizations". In: *Journal of Mathematical Sociology* 24 (1 1999), pp. 67–96.

[69] V. Batagelj and A. Mrvar. "Pajek - Program for Large Network Analysis". In: *Connections* 21 (2 1998), pp. 47–57.

[70] Wouter de Nooy, Andrej Mrvar, and Vladimir Batagelj. *Exploratory Social Network Analysis with Pajek*. 2nd ed. Cambridge University Press, 2011.

[71] Stephen P. Borgatti, Martin G. Everett, and Linton C. Freeman. *UCINET for Windows: Software for Social Network Analysis*. Analytic Technologies, 2002.

[72] G. Csardi. *igraph: Network analysis and visualization.* 2012.

[73] R Development Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: Foundation for Statistical Computing, 2011.

[74] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks". In: 2009. URL: `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`.

[75] M. Smith et al. *NodeXL: a free and open network overview, discovery and exploration add-in for Excel 2007/2010.* Social Media Research Foundation, 2010.

[76] Robert W. Floyd. "Algorithm 97". In: *Communications of the ACM* 5 (1967), p. 345.

[77] Michael L. Fredman and Robert Endre Tarjan. "Fibonacci heaps and their uses in improved optimization algorithms". In: *Journal of the Association for Computing Machinery* 34 (3 1987), pp. 596–615.

[78] Donald B. Johnson. "Efficient algorithms for shortest paths in sparse networks". In: *Journal of the Association for Computing Machinery* 24 (1 1977), pp. 1–13.

[79] Ulrik Brandes. "A faster algorithm for betweenness centrality". In: *Journal of Mathematical Sociology* 25 (2 2001), pp. 163–177.

[80] Jing Yang and Yingwu Che. "Fast computing betweenness centrality with virtual nodes on large sparse networks". In: *PLoS One* 6 (7 2011), e22557.

[81] Saroja Kanchi and David Vineyard. "An optimal distributed algorithm for all-pairs shortest-path". In: *International Journal Information Theories and Applications* 11 (4 2004), pp. 141–146.

[82] Zhiao Shi and Bing Zhang. "Fast network centrality anlaysis using GPUs". In: *BMC Bioinformatics* 12 (2011), pp. 149–155.

[83] Ulrik Brandes and Christian Pich. "Centrality estimation in large networks". In: *International Journal of Bifurcation and Chaos* 17.07 (2007), pp. 2303–2318.

[84] Kazuya Okamoto, Wei Chen, and Xiang-Yang Li. "Frontiers in Algorithmics". In: vol. 5059. Springer-Verlag, 2008. Chap. Ranking of closeness centrality for large-scale social networks, pp. 186–195.

[85] David Eppstein and Joseph Wang. “Fast Approximation of Centrality”. In: 2001.

[86] P. Erdös and A. Rényi. “On the evolution of random graphs”. In: *Publication of the Mathematical Institute of the Hungarian Acadamy of Sciences*. 1960, pp. 17–61.

[87] Albert-László Barabási and Réka Albert. “Emergence of scaling in random networks”. In: *Science* 286.5439 (1999), pp. 509–512.

[88] Albert-László Barabási, Réka Albert, and Hawoong Jeong. “Mean-field theory for scale-free random networks”. In: *Physica A: Statistical Mechanics and its Applications* 272.1 (1999), pp. 173–187.

[89] Sergey N Dorogovtsev, José Fernando F Mendes, and Alexander N Samukhin. “Structure of growing networks with preferential linking”. In: *Physical Review Letters* 85.21 (2000), pp. 4633–4636.

[90] Duncan J. Watts and Steven H. Strogatz. “Collective dynamics of ’small-world’ networks”. In: *Nature* 393 (1998), pp. 440–442.

[91] Yann Bramoullé and Brian W. Rogers. *Diversity and popularity in social networks: Discussion paper*. Tech. rep. 1475. Center for Mathematical Studies in Economics and Management Science, 2009. eprint: `http://hdl.handle.net/10419/31253`.

[92] Mark Bernard and David Seim. “Preferential attachment reloaded: Heterogeneity, multiplicity and a microfoundation”. In: 2009. eprint: `http://www.icfsn.net/docs/bernard.pdf`.

[93] Chunguang Li and Philip K. Maini. “An evolving network model with community structure”. In: *Journal of Physics A: Mathematical and General* 38 (2005), pp. 9741–9749.

[94] Jenny Lennartsson et al. “SpecNet: A spatial network algorithm that generates a wide range of specific structures”. In: *PLoS One* 7 (8 2012), e42679.

[95] Hyoungshick Kim and Ross Anderson. “Temporal node centrality in complex networks”. In: *Physical Review E* (2012), p. 026107.

[96] Thomas W. Valente and Patchareeya Pumpuang. "Identifying opinion leaders to promote behavior change". In: *Health Education & Behavior* 34 (6 2007), pp. 881–896.

[97] C. Christina Mehta. "Centralization in Various Graph Generating Methods: Dissertation Topic 1".

[98] Frank Harary and Edgar M. Palmer. *Graphical Enumeration*. Academic Press, 1973.

[99] Rudolf Mathon. "A note on the graph isomorphism counting problem". In: *Information Processing Letters* 8.3 (1979), pp. 131–132.

[100] Scott Fortin. *The graph isomorphism problem*. Tech. rep. 1996.

[101] Brendan D. McKay and Adolfo Piperno. *nauty and Traces User's Guide v 2.5*. 2013.

[102] Matt J. Keeling and Ken T.D. Eames. "Networks and epidemic models". In: *Journal of the Royal Society of Interface* 2 (2005), pp. 295–307.

[103] G. Witten and G. Poulter. "Simulations of infectious diseases on networks". In: *Computers in Biology and Medicine* 37 (2 2007), pp. 195–205.

[104] Johan Giesecke. *Modern Infectious Disease Epidemiology*. 2nd. Arnold.

[105] Mina Youssef and Caterina Scoglio. "An individual-based approach to SIR epidemics in contact networks". In: *Journal of Theoretical Biology* 283.1 (2011), pp. 136–144.

[106] Alun L. LLoyd and Robert M. May. "Spatial heterogeneity in epidemic models". In: *Journal of Theoretical Biology* 179 (1996), pp. 1–11.

[107] Adam Kleczkowski and Bryan T Grenfell. "Mean-field-type equations for spread of epidemics: The small world model". In: *Physica A: Statistical Mechanics and its Applications* 274.1 (1999), pp. 355–360.

[108] Samuel R. Friedman and Sevgi Aral. "Social networks, risk-potential networks, health, and disease". In: *Journal of Urban Health: Bulletin of the New York Academy of Medicine* 78 (3 2001), pp. 411–418.

[109] Jacco Wallinga, W. John Edmunds, and Mirjam Kretzchmar. "Perspective: Human contact patterns and the spread of airborne infectious diseases". In: *Trends in Microbiology* 7 (9 1999), pp. 372–375.

[110] Michael Boots and Akira Sasaki. "Small worlds and the evolution of virulence: infection occurs locally and at a distance". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 266.1432 (1999), pp. 1933–1938.

[111] Jeffrey O. Kephart and Steve R. White. "Directed-graph epidemiological models of computer viruses". In: *Proceedings from 1991 IEEE Computer Society Symposium on Research in Security and Privacy*. 1991.

[112] Deepayan Chakrabarti et al. "Epidemic thresholds in real networks". In: *ACM Transactions on Information and System Security (TISSEC)* 10.4 (2008), p. 1.

[113] Maksim Kitsak et al. "Identification of influential spreaders in complex networks". In: *Nature Physics* 6 (2010), pp. 888–893.

[114] Frank Bauer and Joseph T. Lizier. "Identifying influential spreaders and efficiently estimating infection numbers in epidemic models: A walk counting approach". In: *EPL* 99 (2012), p. 68007.

[115] John L Wylie et al. "A network view of the transmission of sexually transmitted infections in Manitoba, Canada". In: *Sexually Transmitted Infections* 86.Suppl 3 (2010), pp. iii10–iii16.

[116] Stephen Eubank et al. "Modelling disease outbreaks in realistic urban social networks". In: *Nature* 429.6988 (2004), pp. 180–184.

[117] Chris L Barrett, Stephen G Eubank, and James P Smith. "If smallpox strikes Portland..." In: *Scientific American* 292.3 (2005), pp. 54–61.

[118] Christopher L. Barrett et al. "EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks". In: *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*. IEEE Press. 2008, p. 37.

[119] Mario Ventresca and Dionne Aleman. “Evaluation of strategies to mitigate contagion spread using social network characteristics”. In: *Social Networks* 35.1 (2013), pp. 75–88.

[120] Vittoria Colizza et al. “The role of the airline transportation network in the prediction and predictability of global epidemics”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.7 (2006), pp. 2015–2020.

[121] Linyuan Lü et al. “Leaders in social networks, the delicious case”. In: *PloS One* 6.6 (2011), e21202.

[122] R.M. Christley et al. “Infections in social networks: Using network analysis to identify high-risk individuals”. In: *American Journal of Epidemiology* 162 (10 2005), pp. 1024–1031.

[123] Antonios Garas et al. “Worldwide spreading of economic crisis”. In: *New Journal of Physics* 12.11 (2010), p. 113043.

[124] Nicholas A. Christakis and James H. Fowler. “Social network sensors for early detection of contagious outbreaks”. In: *PloS One* 5.9 (2010), e12948.

[125] Ken T.D. Eames and Matt J. Keeling. “Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases”. In: Proceedings of the National Academy of Sciences of the United States of America (20 2002), pp. 13330–13335.

[126] Andrew Barbour and Denis Mollison. “Epidemics and random graphs”. In: *Stochastic Processses in Epidemic Theory* 86 (1990), pp. 86–89.

[127] Cristopher Moore and M.E.J Newman. “Epidemics and percolation in small-networks”. In: *Physical Review E* 61.5 (2000), pp. 5678–5682.

[128] Romualdo Pastor-Satorras and Alessandro Vespignani. “Epidemic spreading in scale-free networks”. In: *Physical Review Letters* 86 (14 2001), pp. 3200–3203.

[129] Noam Berger et al. “On the spread of viruses on the internet”. In: *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics. 2005, pp. 301–310.

[130] Peter Neal. "SIR epidemics on a Bernoulli random graph". In: *Journal of Applied Probability* 40 (3 2003), pp. 779–783.

[131] P. Schumm et al. "Epidemic speading on weighted contact networks". In: *Bio-Inspired Models of Network, Information and Computing Systems, 2007. Bionetics 2007. 2nd.* IEEE. 2007, pp. 201–208.

[132] Jun-Jun Cheng et al. "An epidemic model of rumor diffusion in online social networks". In: *The European Physical Journal B* 86 (2013), pp. 29–35.

[133] Erick Stattner, Martine Collard, and Nicolas Vidot. "Diffusion in dynamic social networks: Application in epidemiology". In: *Database and Expert Systems Applications*. Springer. 2011, pp. 559–573.

[134] Piet Van Mieghem, Jasmina Omic, and Robert Kooij. "Virus spread in networks". In: *Networking, IEEE/ACM Transactions on* 17.1 (2009), pp. 1–14.

[135] Roy M. Anderson and Robert M. May. *Infectious Diseases of Humans*. Oxford University Press, 1991.

[136] O Diekmann, JAP Heesterbeek, and Johan AJ Metz. "On the definition and the computation of the basic reproduction ratio R 0 in models for infectious diseases in heterogeneous populations". In: *Journal of Mathematical Biology* 28.4 (1990), pp. 365–382.

[137] M Elizabeth Halloran. "Concepts of transmission and dynamics". In: *Epidemiologic methods for the study of infectious diseases* (2001), pp. 56–85.

[138] Romulus Breban, Raffaele Varavas, and Sally Blower. "Theory versus data: How to calculate $R_0$". In: *PLoS One* (3 2007), e282.

[139] Annett Nold. "Heterogeneity in disease-transmission modeling". In: *Mathematical Biosciences* 52.3 (1980), pp. 227–240.

[140] Yang Wang et al. *Epidemic spreading in real networks: An eigenvalue viewpoint*. Tech. rep. 544. Computer Science Department, 2003.

[141] Ayalvadi Ganesh, Laurent Massoulié, and Don Towsley. "The effect of network topology on the spread of epidemics". In: *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*. Vol. 2. IEEE. 2005, pp. 1455–1466.

[142] Thealexa Becker et al. "Virus dynamics on starlike graphs". In: *arXiv preprint arXiv:1111.0531* (2011).

[143] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[144] Piet Van Mieghem. *Graph Spectra for Complex Networks*. Cambridge University Press, 2011.

[145] E. R. van Dam and R. E. Kooij. "The minimal spectral radius of graphs with a given diameter". In: *Linear Algebra and its Applications* 423 (2007), pp. 408–419.

[146] Jun Yan. "geepack: Yet Another Package for Generalized Estimating Equations". In: *R-News* 2/3 (2002), pp. 12–14.

[147] Jun Yan and Jason P. Fine. "Estimating Equations for Association Structures". In: *Statistics in Medicine* 23 (2004), pp. 859–880.

[148] Sren Hjsgaard, Ulrich Halekoh, and Jun Yan. "The R Package geepack for Generalized Estimating Equations". In: *Journal of Statistical Software* 15/2 (2006), pp. 1–11.

[149] C. Christina Mehta. "Various Properties of Centralization in Small Graphs: Dissertation Topic 2".