Network Reliability: Theory, Estimation, and Applications

Yasamin Khorramzadeh

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Physics

Stephen Eubank, Co-Chair
Uwe C. Täuber, Co-Chair
Vito Scarola
Michel Pleimling
Jean J. Heremans

December, 2015
Blacksburg, Virginia

Keywords: Complex Networks, Network Reliability, Network Topology

Network Reliability: Theory, Estimation, and Applications

Yasamin Khorramzadeh

(ABSTRACT)

Network reliability is the probabilistic measure that determines whether a network remains functional when its elements fail at random. Definition of functionality varies depending on the problem of interest, thus network reliability has much potential as a unifying framework to study a broad range of problems arising in complex network contexts. However, since its introduction in the 1950's, network reliability has remained more of an interesting theoretical construct than a practical tool. In large part, this is due to well-established complexity costs for both its evaluation and approximation, which has led to the classification of network reliability as a NP-Hard problem. In this dissertation we present an algorithm to estimate network reliability and then utilize it to evaluate the reliability of large networks under various descriptions of functionality.

The primary goal of this dissertation is to pose network reliability as a general scheme that provides a practical and efficiently computable observable to distinguish different networks. Employing this concept, we are able to demonstrate how local structural changes can impose global consequences. We further use network reliability to assess the most critical network entities which ensure a network's reliability. We investigate each of these aspects of reliability by demonstrating some example applications.

Acknowledgments

I would like to thank the many people who made this research a reality. I would like to thank my co-chairs, Dr. Stephen Eubanks, for his knowledge, support and his essential contributions to this work, and Dr. Uwe C. Täuber for his guidance and advising efforts from the beginning to the end of my doctoral studies. Also, I extend many thanks to Dr. Vito Scarola, who I learned a lot from during the early years of my doctoral studies. Thank you to the other members of my thesis committee, Dr Michel Pleimling and Jean J. Heremans, for their great feedback on this thesis and the inspiring courses. I am extremely grateful to Dr. Mina Youssef and other colleagues who collaborated with me on this work over the years. I would also like to take the chance to thank Dr. Samarth Swarup for his generous guidance and practical advice during the last year of my studies. Further appreciation goes out to the faculty, staff, and fellow graduate students at the Network Dynamics and Simulation Science Lab whose daily support and assistance made the whole process more interesting and rewardingno one more so than Dr. Madhav Marathe for his great leadership and constant encouragement.

Further appreciation goes out to the faculty, staff, and fellow graduate students at the Network Dynamics and Simulation Science Lab whose daily support and assistance made the whole process more interesting and rewardingno one more so than Dr. Madhav Marathe for his great leadership and encouragement.

It is hard to overstate my gratitude to Trevor Flanery for being a source of strength and support during the final stages of this thesis. I would also like to warmly thank my family and friends whose emotional support spanned oceans. Lastly, and most importantly, I wish to thank my parents, Zoleikha Darvish-Basseri and Zal Khorramzadeh for their endless love, support and patience through all of my studies.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

Chapter I

Introduction

Complex networks have come to play an increasingly important role in diverse fields of study, from biology to engineering to social sciences. Network science utilizes concepts from graph theory to model systems of interest. A primary area of network study concerns how dynamical processes diffuse across networks. These diffusive processes include energy [57], information, or commodity exchanges from chemical atoms [70] to global pandemics [60]. We use network reliability to study these diffusive processes on networks and different research problems emerging around this topic.

For instance, in social science a population can be modeled as a network consisting of nodes, each representative of an individual in the population, and edges denoting relationships or flows of some commodity between the nodes. An infectious disease spreading through a contact network can be modeled as a diffusive process where individuals come into contact in ways that support disease transmission. In such studies finding the transmission tree–the movement of infection from node to node–is a recurring research problem. Estimating the transmission tree sheds lights on the dynamics of the disease through the population and gives insights into the mechanism of transmission.

Network reliability combines the structural details of the network with the dynamical processes and thus provides a useful tool to investigate diffusion related problems.

Major areas of research regarding diffusive phenomena on networks can be classified into three different categories. Firstly, a set of problems investigates the influence of *structural properties* on the diffusion outcome. For example, these studies allow researchers to investigate how peoples' interactions in a population affect important features of the disease transmission.

The second set of problems is the associated issues of vulnerability and controllability. Vulnerability research explores interventions and control strategies by assessing the set of *central* network elements, which, if properly controlled, can suffice to achieve desired outcomes. Examples of this research include studies aimed to unravel the most important people in the population to vaccinate to significantly decrease the outbreak size of an epidemic. These studies are particularly important, since many networks have expanded over large geographical or ethnographical spaces, making them impossible to entirely control during outbreaks. Therefore, assessing the set of most important network entities allows us to better control the system given limited resources.

Finally, numerous generating network models are proposed to estimate real world networks. To vet a model it must be compared to existing network data, and thus the outcomes of diffusive processes can provide a useful *observable* for such validation. Were we to generate network models that demonstrate similar epidemic spread to actual epidemics events, network reliability can provide a measure of the outcome of the disease outbreak.

To summarize, complex networks have been used to model the patterns of interaction of different systems. While specific dynamics and system requirements vary, some sets of emerging problems overlap between different disciplines. Thus, it is crucial to develop common frameworks that can capture features of the systems of interest and provide tools for comprehensively assessing them. The primary focus of this

dissertation is to expand network reliability as such a tool.

Portions of this chapter were originaly published in the Journal of Complex Networks, in 2014, in volume 2, issue 4, pages 356-372. [27], the Journal of Physical Review E, in 2013, volume 88, issue 5 [79] and the Journal of Physical Review E, in 2015, volume 91, issue 4 [42].

1.1 Foundations of Network Reliability

Shannon and Moore [48] first introduced network reliability in their 1950's study of the reliability of relay circuits. They posed network reliability as the probability that a circuit remains closed, given that the individual contacts comprising it are closed with some known probability. In the more than half century since their work was first published, a large body of research has developed around this topic. In various disciplines this same concept has also been described by such terms as performability [20], survivability [11] and performance [68].

However, owing to its computational complexity, network reliability has mostly remained as a theoretical concept. In this project we propose a Monte Carlo algorithm that provides an estimation of the reliability within desired accuracy. This simulation method assists us to apply this concept to large real world network problems. We explore some of the potential applications in later sections.

This chapter provides an overview of network reliability. Section 1.2 provides a description of the concept, including definition of the network model. We provide general definitions to emphasize the broad applicability of the formalism and methods we will develop. Section 1.3 surveys various representations of network reliability and their interpretation. The computational complexities and our proposed algorithm are described in Section 1.4.

1.2 Network reliability basics

Reliability is a probabilistic measure of diffusive processes on networks. Colbourn [18] provides a concise description of network reliability which encapsulates other definitions. According to Colbourn network reliability is "the probability that a network remains functional in the face of random failures of its elements."

This definition is comprised of three different components:

1. **Network:** The abstraction of the system under study as a graph is referred to by a network denoted by $G(V, E)$, where V is the number of nodes (vertices), and E represents the number of edges between them. Complex networks play an increasingly important role across different disciplines. For instance, in epidemiology the fact that disease can be transmitted from one individual to another via contact, when paired with the observation that each person has a limited set of contacts, can be best represented by a network where nodes represent different individuals and edges are the contacts between them. A different example is the relay network in Shannon and Moore's original study of reliability, where each contact was represented as a node in the network and wires connecting the contacts are links in the graph. A third example is a radio broadcasting network [1] where each node represents a station and an edge connects two sites only if the corresponding stations can directly communicate with each other.

   Different network models can be adopted to incorporate the topological information of a given system. Basic network models often assume undirected simple graphs, as in the three examples described above. This general network model allows more variation, such as directed edges representing directional links in digraph models, or permitting multiple links between nodes in multigraph mod-

els.

2. **Element's failure:** Network elements (nodes and/or edges) are subject to failure, which can be modeled by what is known as a failure model. A failure model associates with each element i of the network a probability p_i that the element operates and the complementary probability $q_i = 1 - p_i$ that it fails. Each element operates or fails independently from other elements in the network. One simplifying assumption is for all elements to have the same probability of failure If that is the case, the element subscripts are not necessary.

   Though the general network reliability definition allows for both node and edge failures, more often the study of reliability utilizes failure models that consider one set of elements to be completely operational while the other set's functionality is probabilistic. The most common model of this type is known an **edge failure damage model**, with nodes that operate perfectly and edges which fail with some probability. This damage model corresponds to the **bond percolation model** where each edge is open, independent of other edges with a probability p and its close otherwise. Consequently the reverse case corresponding to the **site percolation model** is called the **node failure damage model** and has been of a great interest. A more complex scenario where failure of both the edges and nodes is allowed, is less commonly used.

   For a network used in the study of epidemiology, an edge failure damage model, describes the probabilistic nature of disease transmission. In other words, not every contact between an infected individual and a susceptible person leads to disease transmission. The transmission happens with some probability dependent on the duration and type of contacts, along with other factors. It is intuitive that not everyone contracts diseases with the same probability, i.e., kids and elderly are known to be more vulnerable in many cases. This observa-

tion can be described by a node failure model where a probability is assigned to each individual's vulnerability. The most general case, where both nodes and edges fail, can model the probabilistic transmission of a disease via contacts alongside individual vulnerabilities.

Returning to the Moore and Shannon circuit network, assuming that the links cannot fail but each contact could be closed with some probability, the equivalent damage model is a node failure model. Finally, employment of the simultaneous nodes and edges failure model could be used to model the radio broadcasting network. There, nodes could be in operational state or not, and links between two sites can potentially fade with some probability due to natural conditions that may cause spontaneous or sustained damages.

More importantly, a node failure model can easily be converted to an edge failure one while preserving reliability. For this transformation a node is replaced by two perfectly reliable nodes connected by an edge with failure probability equal to that of the initial imperfect node. Therefore, study of the reliability of networks using edge failure model can encompass both edge and node failure cases, making it more desirable to work with this type of model.

3. **Functional:** The most important feature of network reliability lies in the description of functionality. What does it mean for a network to be functional? The answer is a problem-specific description which has broadened applications of network reliability to a wide range of problems. A network is deemed functional if it has a specific property related to its intended use. For instance, in the case of Shannon and Moore's switched circuits the property is for the circuit to be closed between two pre-specified points.

   To assess whether a network is operational or not, a binary function ϕ is defined which applied to graph $G(V, E)$, returns 1 or 0 indicating whether it has

the desired property or not, i.e. $\phi : \mathcal{G} \rightarrow \{0, 1\}$ where $\mathcal{G}$ is the set of all subgraphs of G. A well-studied property is two-terminal reliability that defines a network as operational only if there exists a path between two specified nodes, commonly called source and terminal nodes. Direct practical application of this rule emerges in the study of radio broadcasting networks, as reliability in this case characterizes the ability of two nodes to communicate. In other words the property of interest in this case, not so distant from Shannoon and Moore's circuit property, is whether there exists a connected path between the two points. In this case, reliability answers this question: Beginning from a source station, can our message be delivered to a desired site by relaying, through one or more other stations, even in the absence of a direct link?

A second commonly studied property is the all-terminal property [18, 17, 58], which also appears in network broadcasting studies. This property can be considered as a generalization of the two-terminal case, where each node is presumed to be a terminal node, thus the network is considered to be operational if for every pair of nodes in the graph there exists a path between them.

Finally there is the intermediate case known as K-terminal reliability [58, 66] which defines a network as operational when there exists a path between each pair of nodes in a set of K predefined vertices. To illustrate, consider for instance a distributed processing system with a set of K sites that can either collectively run a task or provide resources required for other nodes. Implementing the system as a network with unreliable nodes, we are interested in obtaining a network where each operational pair of sites in the set of K nodes can communicate with one another.

In addition to the properties described here, one can potentially define other properties to identify an operational state if one can frame the condition under study as a binary function. Therefore, network reliability appears to be a

suitable tool for the study of complex networks emergent in different problems.

1.3 Mathematical representation of Network Reliability

Consider a graph $G(V, E)$ with V perfectly reliable nodes and E unreliable edges. For simplicity, assume the edges to be undirected and to have a uniform probability of failure of $1 - x$, i.e., an edge failure model is assumed. The extension to directed edges is immediate; but it is slightly more complicated to account for a heterogeneous failure probability. Let g be a subgraph of the network: $g \subseteq G$. The damage model D_x assigns the probability that this subgraph occurs. Using the edge failure damage model, it is equivalent to the probability that all the edges comprising the subgraph are in an operational state and all remaining edges in the network are not. Now whether subgraph g is operational or not can be indicated by applying the property function $\phi(g)$ which returns 1 or 0 respectively.

Reliability is the expected value of the property function ϕ over all possible subgraphs weighted by the probability of the subgraph under the damage model:

$$R_G(D_x, \phi) \equiv \sum_{g \subset G} \phi(g) D_x(g) \tag{1.1}$$

Therefore, reliability is the measure of expected performance in the face of element failures. In most cases, however, for a specific study, one is concerned about a particular performance measure and thus identifies a specific property function as well as a distinct damage model. In this case we leave out the explicit dependence on D and ϕ yielding:

$$R(x) = \sum_{g \subset G} \phi(g) D_x(g) \tag{1.2}$$

Since the property function $D_x(g)$ has a binary form, the summation in Equation 1.2 reduces to the sum over the functional subgraphs:

$$R(x) = \sum_{g \subseteq G_{Functional}} D(g) \tag{1.3}$$

where sum is over the operational subgraphs only ($G_{Functional}$).

It is often the case that many subgraphs have the same probability of occurrence under a damage model. We can rewrite the reliability in terms of the set of equivalence classes induced by the damage model. Thus, if there are C such classes $\{c_1, \ldots, c_C\}$ and the ith class contains $|c_i|$ subgraphs, each appearing with probability $p_{\vec{x}}(c_i)$, then:

$$R(\vec{x}) = \sum_{k=1}^{C} \sum_{g \in c_k} \phi(g) D_x(g) \tag{1.4}$$

$$= \sum_{k=1}^{C} R_k D_x(c_k) \tag{1.5}$$

where R_k represents the number of subgraphs in class k that passed the functionality test, i.e. $\phi(g_k) = 1$. Assuming an edge failure damage model, equivalence classes under this damage model are sets of subgraphs that have the same number of edges. There are $C = E + 1$ such sets. The probability of selecting any particular subgraph of size k is simply the product of the probability of selecting each of its edges while not selecting any of the rest of the edges that do not appear in that subgraph, reducing it to a particularly simple form: $D_x(g_k) = x^k(1-x)^{E-k}$, where g_k denotes the subgraph

with k edges. Substituting in Equation 1.4, we obtain:

$$R(x) = \sum_{k=0}^{E} R_k x^k (1-x)^{E-k} \tag{1.6}$$

Here R_k denotes the number of reliable subgraphs with k operating edges.

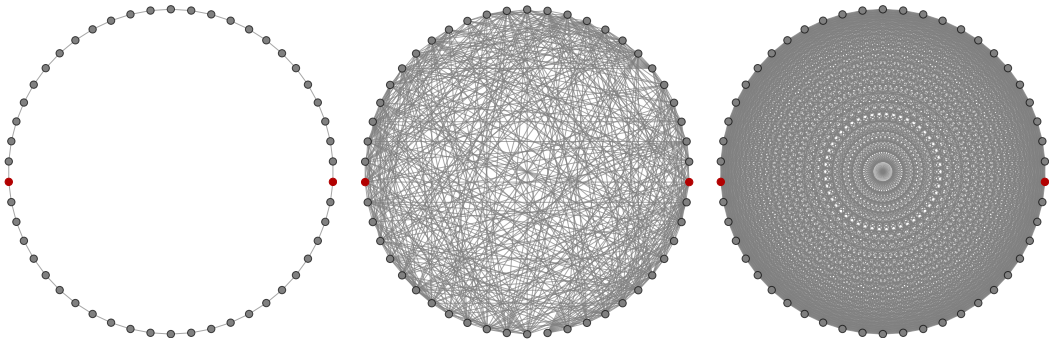


Figure 1.1: Three possible networks connecting $N = 50$ nodes: (left) a circle network; (middle) a graph in which each possible edge is present with probability 0.4, GNP(50,0.4); (right) a Complete network, K50, in which every possible edge is present. Nodes 1 and 25 in each network are colored red.

This notation leads to a polynomial of degree at most E in one variable, the probability of edge reliability x, known as the **reliability polynomial**. $R(x)$ is a monotonically non-decreasing, continuous function passing through the origin and point $(1, 1)$. Figure 1.2 illustrates the reliability of networks depicted in Figure 1.1 under the two-terminal reliability rules. It can be seen that reliability depends on the topology of the network. It is intuitive that in the complete graph there are more paths available to connect source and target nodes than those that are available in a random and cycle network respectively. The only zero of the reliability occurs at the origin; however, there is typically a narrow range of x in which it increases from near 0 to near 1, and the width of this range shrinks as the number of edges, E, increases.

This representation of reliability is particularly important in network design problems, when the goal is to characterize the network structure in a manner relevant to

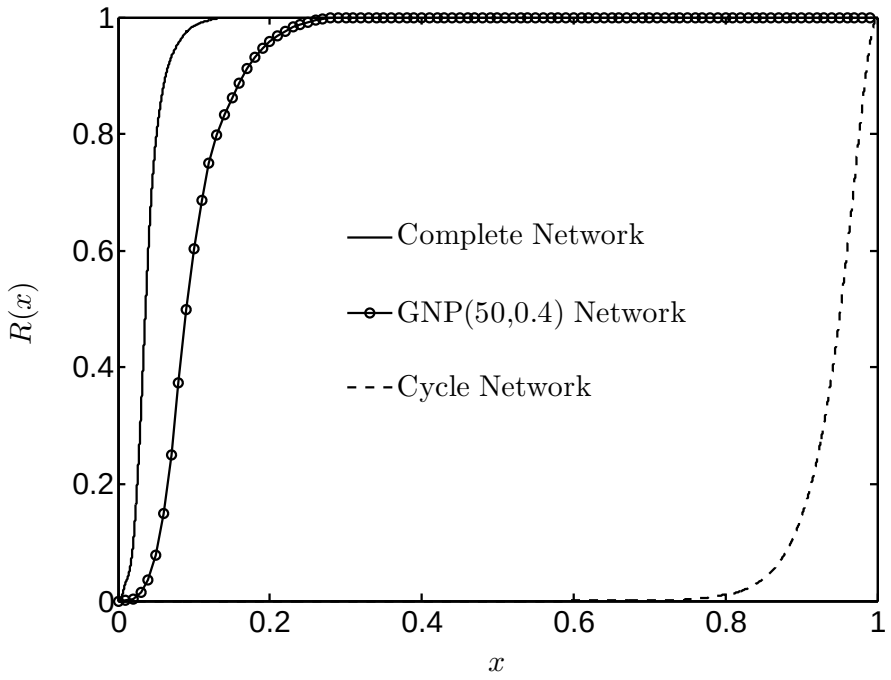


Figure 1.2: Reliability polynomial for Complete, GNP and Circle graphs as depicted in Figure 1.1. Reliability criterion is the two-terminal property with nodes 1 and 25 as source and target nodes respectively (colored red). The complete graph is the most reliable, meaning that it is the most probable to have a path between source and target nodes, due to numerous redundant paths. The cycle graph is the least reliable configuration since there is only one possible path between the source and target nodes. Finally, the GNP graph is an intermediate structure where a subset of redundant edges, and consequently the existence of more paths, make it more reliable than the Cycle graph but less reliable than the Complete graph.

the dynamical process taking place upon it. For instance, in network epidemiology we are interested in designing interventions that make the network less reliable when studying an outbreak. In this case, different intervention strategies require different structural changes to the initial network. Comparing the reliability polynomials of these proposed structures identifies which modification builds the less reliable structure. An interesting observation first reported by Moore and Shannon [48] is the crossing of reliability curves, implying that for a range of x values one configuration of edges might be more reliable than the other, while for the remainder of the range the reverse holds true. This observation emphasizes that network reliability, in addition to the structural properties of the network, depends on the edge failure probability. This poses network reliability as an observable providing us a quantitative way to compare graph structures with respect to specific dynamics. We investigate this problem further in Chapter III.

1.3.1 Reliability representation

We know that there are $\binom{E}{k}$ subgraphs with k edges in the graph, and that some fraction between 0 and 1 of them have the property of interest. Therefore, the reliability coefficient can also be written as follows:

$$P_k \equiv R_k / \binom{E}{k}. \tag{1.7}$$

where P_k is the fraction of subgraphs with k edges that have the property ϕ. Hence:

$$R(x) = \sum_{k=0}^{k=E} P_k \binom{E}{k} x^k (1-x)^{E-k}. \tag{1.8}$$

This decomposition splits R_k into what we call an *entropic* or combinatorial factor $\binom{E}{k}$

and a *structural* factor P_k. The entropic factor makes explicit the sharp peak in the number of possible subgraphs with k edges, i.e. the size of the space from which equiprobable system configurations can be drawn. It creates an envelope that windows the effects of $x^k(1-x)^{E-k}$ on a small region centered at $k = Ex$. The factor P_k is structural in the sense that it encodes all the information about the specific graph G that is needed to determine its reliability. This is obvious because the probability $x^k(1-x)^{E-k}$ and the entropic factor $\binom{E}{k}$ are the same for any graph with E edges.

Note that although this is a continuous polynomial in the variable x, it involves only a finite number of coefficients $R_k, k \in \{0, \ldots, E\}$. That is, $R(x)$ is an element of an $E+1$-dimensional vector space with basis elements $x^k(1-x)^{E-k}$. There are, of course, many other bases we could choose for the reliability polynomial. An orthogonal basis, such as the first $E+1$ Legendre polynomials, might have useful estimation properties. Here we use another non-orthogonal basis–the functions x^k–because of its simplicity and attractive interpretation. There is a unique mapping from coefficients in one basis to those in the other, which can be derived by expanding the factor $(1-x)^{E-k}$ in Equation 1.6:

$$\begin{aligned}
R(x) &= \sum_{l=0}^{E} R_l x^l \sum_{m=0}^{E-l} \binom{E-l}{m} (-1)^m x^m \\
&= \sum_{l=0}^{E} R_l \sum_{l+m=l}^{E} \binom{E-l}{m} (-1)^m x^{l+m} \\
&= \sum_{l=0}^{E} R_l \sum_{k=l}^{E} \binom{E-l}{k-l} (-1)^{k-l} x^k \\
&= \sum_{k=0}^{E} (-1)^k x^k \sum_{l=0}^{k} (-1)^l R_l \binom{E-l}{k-l} \\
&= \sum_{k=0}^{E} N_k x^k
\end{aligned} \tag{1.9}$$

where

$$N_k \equiv (-1)^k \sum_{l=0}^{k} (-1)^l \binom{E-l}{k-l} R_l. \tag{1.10}$$

The former basis has an interesting interpretation in terms of what we call *structural motifs*. These are structural building blocks that play an important role in the dynamical process. Structural motifs support analytical reasoning about the consequences of changes in network structure. We explore this in depth in Chapter III of this manuscript.

1.4 Computational complexity

Reliability provides a natural way to characterize a graph that is immediately related to the dynamic phenomena taking place on it. However, exact determination of coefficients in the reliability polynomial for many rules and many classes of networks is known to be intractable [19, 5, 18].

As can be seen in Equation 1.1, reliability depends on the probability of occurrence for every subgraph of graph G under the chosen failure model, denoted by the $D(g)$. For instance, as described above, the independence edge failure model, simplifies the computation of $D(g)$ to the probability with which any subgraph of size k occurs, i.e.: $D(g_k) = x^k(1-x)^{E-k}$

Moreover, for many measures of functionality including two-terminal, K-terminal and all-terminal reliability, the computation of $\phi(g)$ is also straightforward. For two-terminal reliability this is done by finding a path between the source and target nodes whereas, for all-terminal reliability the problem is equivalent to finding a tree structure expanding the network. Nevertheless, the evaluation of 1.1 remains intractable; indeed exact evaluation appears to require exponential time. This can be explained by a closer look at the mathematical definition of network reliability.

Computing reliability requires what is known as the *exact state enumeration* that runs over the set of all possible subgraphs of G, and for each subgraph determines whether it is operational or not. The computational problem arises due to the enormous number of subgraphs involved and the exponential growth of this number with the size of the network. In other words, for a graph $G(V, E)$ with E unreliable edges, there exist 2^E subgraphs. Therefore an exact technique cannot deal with a network of even 100 elements. In general, it has been shown that it is unlikely to find a polynomially bounded algorithm able to compute network reliability. Thus, network reliability problems have been classified as NP-complete problems [19, 5, 18], meaning that it is unlikely to find polynomial time solutions for them.

To cope with the complexity of computing reliability, a number of directions have been explored. One avenue of research has focused on ways to take advantage of particular properties or classes of graphs. [67, 2, 41, 74]

A different approach has been constructed around estimating the network reliability where computation is restricted to only a sample of the state space. This approach has led to a series of *Monte Carlo* algorithms which appear to provide efficient approximation for different reliability measures. We, too, use a Monte Carlo method to estimate network reliability.

Our estimation procedure is simply to evaluate the reliability of each of a sample of n subgraphs drawn from G in the same equivalence class under the damage model. For example, under an independent edge damage model, we draw a sample of subgraphs that include exactly k edges selected uniformly independently at random from G. The resulting estimate of $\hat{P}_k$ is the fraction of subgraphs in the sample that are accepted:

$$\hat{P}_k = \frac{\sum_{g \in \mathcal{S}} \phi(g)}{n} \tag{1.11}$$

where $\mathcal{S}$ is the sample of n subgraph.

We expect the number of accepted subgraphs in our sample to be distributed binomially with parameters nP_k and n. Then $\hat{P}_k$ is an unbiased estimator for P_k with variance $\frac{\hat{P}_k(1-\hat{P}_k)}{n}$. Thus the estimation precision depends on the sample size and not the space of the subgraphs.

Furthermore, the estimation procedure is an embarrassingly parallel computation, since the sampling and rule evaluation can be performed independently and asynchronously. Finally, to demonstrate that these techniques are feasible for large graphs, we use an estimated social contact network for the New River Valley region near Blacksburg, Virginia. This network was generated by the Network Dynamics and Simulation Sciences Laboratory according to methods described elsewhere, and is publicly available [72, 51]. It contains more than 4.1 million edges and 150 thousand vertices. For this network, we used a different reliability property which we call AR-α-reliability. This property identifies a graph as reliable if α fraction of nodes are in the subgraph. We estimated reliability for three different values of α: 20%, 50% and 80%. As can be seen in Figure 1.3 the network is less reliable if the rule requires more fraction of nodes to be contained in accepted subgraphs.

1.5 Conclusion

The classical concept of network reliability provides a rich theoretical basis, supported by computational estimation procedures, for the study of diffusive processes on networks. Although an exact evaluation of the reliability polynomial is usually intractable, our method can estimate the coefficients within any desired confidence interval. This dissertation aims to reduce the gap between theoretical analysis of reliability problems and our ability to apply the conceptual framework to practical

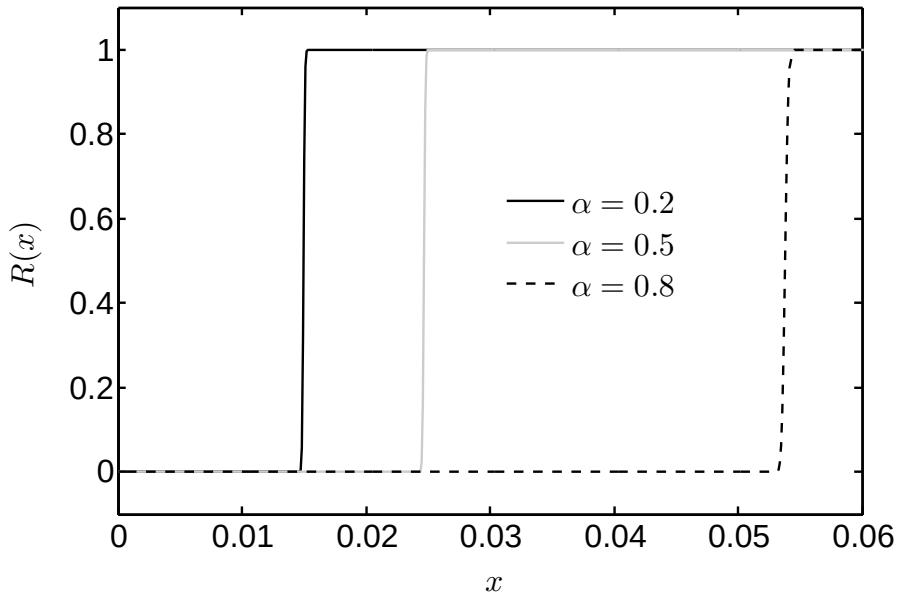


Figure 1.3: Network reliability for New River Valley network under reliability property AR-α-reliability with $\alpha = 20\%$, $\alpha = 50\%$ and $\alpha = 80\%$.

problems for large and non-trivial graphs.

The rest of this dissertation is organized as follows; first we explain in chapter II how we employed reliability to assess the impact of local structural properties on the outcome of the diffusive process on network. A different representation of the reliability is illustrated in Chapter III and further employed to assess dynamically important features between different networks. We further build on this representation to develop a centrality measure and explore its performance against several canonical centrality measures from the literature, as discussed in Chapter IV. Finally, in Chapter V we discuss how network reliability could be used to characterize network structure in a manner directly related to dynamics taking place upon it.

Chapter II

Network Reliability: The effect of local network structure on diffusive processes

2.1 Abstract

The contact network analysis enables us to study how people interact in social structures and thus offers important benefits for the assessment of risk factors and conditions contributing to the disease transmission. A large body of studies explored the network structure and how it influences the spread of epidemic. The aim of these studies is to characterize different aspects of individual properties and local patterns of the contact network, and determine how specific aspects of contact patterns can alter the spread of infectious diseases. For instance, Marathe et al., [46] studied the impact of the mixing pattern at household level on the population level attack rate. By changing the household contact pattern from a fully mixed network to a network with selective contacts, they showed that the global attack rate reduces significantly for a wide range of the transmission rates.

However, our knowledge of the details of the contact network is very scarce, specially in regards to larger networks. Thus, finding the set of parameters that faithfully characterize the intricate structure of contact network has been of a great interest. Wallinga et al., suggested [75] the average number of contacts, the network's transitivity and the characteristic path length as structural metrics that have an important bearing on the spread of infectious diseases.

Moreno, et. al., [50] showed that variation of the node connectivity distribution increase the number of infection's incidence, demonstrating that average topological properties—mean node connectivity in this case—fail to fully capture the overall behavior of epidemic spreading.

Furthering in the same direction, in this paper, we illustrate how to use the concept of network reliability to elucidate how details of network topology influence the spread of epidemics. There are many structural aspects of contact networks that interact in complicated ways with each other and with the dynamical properties of disease transmission to create population-level dynamics in infectious disease outbreaks. Here we explore the impact of local structural changes—particularly node degree asortativity and the clustering coefficient—on network reliability. Despite other studies that are constrained to use summary statistics such as degree distribution or clustering coefficient, as graph representation, network reliability works with the entire adjacency matrix of the network. For a simple undirected graph, its adjacency matrix A attributes 1 to element $A_{i,j}$ if nodes i and j are adjacent (connected via an edge) and otherwise $A_{i,j} = 0$, thus matrix A has $2E$ non-zero entries. Therefore, network's adjacency matrix and consequently the network reliability which uses adjacency matrix as graph representation, contain the accurate and complete description of the contact pattern.

The outline of this study is as follows: First, we re-introduce network reliability in

terms of a reliability property relevant to network epidemiology. Then we estimate reliability for an *in silico* laboratory of networks with a range of carefully controlled topological properties. We characterize these networks' reliability in terms of critical points and other features, elucidating the relationship between network reliability and common graph statistics as a function of network size. Finally, we indicate some intriguing open research problems. I emphasize my contribution to this study; detailed description of the study is presented elsewhere [79].

2.2 Network reliability to study epidemiology

Consider a graph $G(V, E)$ with V vertices and E weighted edges. Let the set $\mathcal{S}$ be the set of all subgraphs of G generated by including each edge (i, j) independently with probability $x_{i,j}$.

Now consider a binary function $\phi : \mathcal{G} \to \{0, 1\}$, which is a quantitative indicator of the property of interest. If $\phi(g) = 1$, we say that subgraph g is *reliable* or *accepted*. We define the *reliability* $R(G, \phi, \vec{x})$ of a base graph G with respect to the reliability property ϕ and edge reliability probability $\vec{x}$ as the probability that a randomly chosen subgraph g is reliable. In other words, a network is reliable to the extent that it remains functional under random removal of edges. Formally:

$$R(G, \phi, D_{\vec{x}}) \equiv \sum_{g \in \mathcal{G}} \phi(g) D_{\vec{x}}(g). \tag{2.1}$$

We will explicitly include the dependence on the graph G and the property ϕ in notation such as $R(G, \phi, D_{\vec{x}})$ only when we wish to distinguish the reliability of two different graphs or two different properties. See Colbourn [18] for a comprehensive introduction to notions of reliability.

To apply reliability to the context of network epidemiology, we have to redefine what

each of the terms in this definition means. Network, clearly, is the representation of the population as a graph, where individuals are nodes and edges represent contacts between them. Hence, the big advantage of using network reliability is that it makes no assumption about the contact pattern, in contrast it includes all the structural information regarding contact network in form of network's adjacency matrix. An infected individual can infect his neighbors with some probability depending on the infectiousness of the disease. Associating a probability x to each edge is thus akin to defining an edge failure model, where probability x that an edge operates represents the disease transmission probability. From a different perspective, one can assume that each individual may contract the disease with some probability if they are in contact with an infected person. In this approach we can assign a probability x to each node, representing the susceptibility of the individual. This can be categorized as the node failure damage model. The more comprehensive model permits both a probabilistic transmission over contacts and a probabilistic susceptibility with respect to the disease. This approach can be modeled by the simultaneous node and edge failure model. Whether the network is functional or not relies on the specific property of interest. *Attack rate (AR)-α* indicates a subgraph is reliable if the mean component size across all *vertices* is greater than or equal to αV. We primarily use the AR-α rule in this paper because of its epidemiological relevance. As its name suggests, it gives the probability that the cumulative fraction of vertices infected (sometimes called the "wet set" in a percolation setting) exceeds α, averaged over all possible initial conditions in which a single vertex is infected. This property, along with many other commonly used reliability properties, has the useful property of *coherence*, i.e. adding an edge to a reliable subgraph does not make it unreliable.

To explore the impact of the network topology on the overall behavior of epidemic, we inspected the reliability, as defined, for two ensembles of graphs, with similar degree distribution and a range of carefully controlled topological properties. The procedure

is explained briefly in the next section.

2.3 Numerical evaluation

Traditionally, Erdős-Rényi (ER) random graphs are used to describe the complex topology of networks. This model is denoted with its two parameters : the number of nodes and a probability that an edge exists between each pair of nodes. Subsequently, the degree distribution of ER graphs follow a binomial distribution. However, recent investigation shows that most real world networks from human contact networks to World Wide Web and power-grid networks, exhibit a power-law degree distribution, i.e. the probability that any node is connected to k other nodes is proportional to $k^{-\gamma}$ where the power law exponent γ is known to be between 2 and 3 for most network. Therefore, the highly heterogeneous class of networks called scale-free networks has been proposed to provide a more acute description of real world networks. Motivated by this observation, in this paper we provide a detailed numerical study of the network reliability on these two classes of networks.

Starting with a randomly generated network of each class, we construct an in silico laboratory of graphs of the same degree distribution, and different local structural properties, here we use degree-assortativity and number of triangles that are explained in the next section. Both classes have the same number of edges $E = 992$ and same number of vertices $V = 341$ thus similar mean degree, the degree distribution for each class is demonstrated in Figure 2.1.

2.3.1 Degree assortativity

Assortativity is the tendency of the nodes to connect preferentially to other nodes with similar attributes. In the case of the degree assortativity this attribute is the

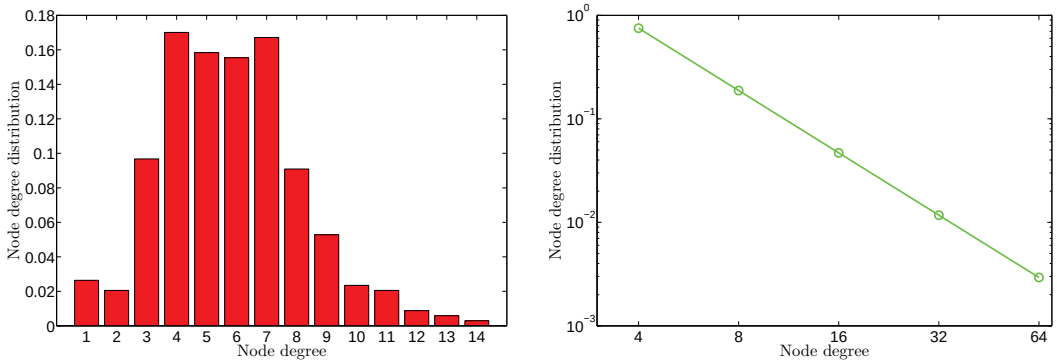


Figure 2.1: Degree distributions for the Erdős-Rényi graphs, ER (left panel); and the scale-free graphs, SFL (right panel). Note the logarithimic axes for SFL graph.

degree of the nodes. It is mathematically defined as the Pearson correlation coefficient of degree of nodes at either ends of an edge.

$$r = \frac{\sum_{xy} (e_{xy} - a_x b_y)}{\sigma_x \sigma_y} \tag{2.2}$$

where e_{xy} denotes the fraction of all links that connect the nodes with degree x and y. a_x and b_y represent the fraction of links that start and end at nodes with degree x and y. σ_x denotes the standard deviation of the variable x. Since r is a correlation coefficient, its value is in the range of $[-1, 1]$ where negative and positive r represent disassortative mixing and assortative mixing respectively. It has been shown that social networks are generally more assortatively mixed, while technological and biological networks seem to have disassortative mixings.

2.3.2 Number of triangle

In many networks, if node A is connected to both nodes B and C, then it is highly probable that B is also directly connected to C. This characteristic is usually quantified by clustering coefficient of a node, the number of triangles a node is a corner

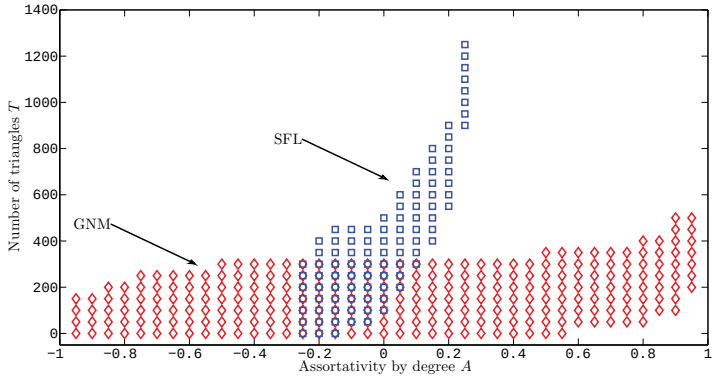


Figure 2.2: Accessible ranges of assortativity by degree and number of triangles for GNM and SFL graphs.

of, normalized by the total number of triangles that could pass through that node. Average of the node clustering over all nodes in the network, is known as the average clustering coefficient of the network and has been the subject of several studies.[77, 7]

$$C = \sum_{l \in V} \frac{2n_l}{k_l(k_l - 1)} \tag{2.3}$$

Where n_l is the number of triangles passing through node l. The big advantage of using the normalized form is that the resulting coefficient is in range of $[0, 1]$ thus providing an intuition of how locally clustered the network is. However, for the sake of simplicity, in preparing the networks' samples in this study, we merely used the total number of triangles in the graph, denoted by T.

Applying an edge shuffling method that preserves the degree distribution while modifying assortativity and number of triangles, we constructed two sets of networks where assortativity-by-degree and number of triangles are expanded over a range, as depicted in Figure 2.2. Network preparation performed by coauthors is explained in depth in the original manuscript [79].

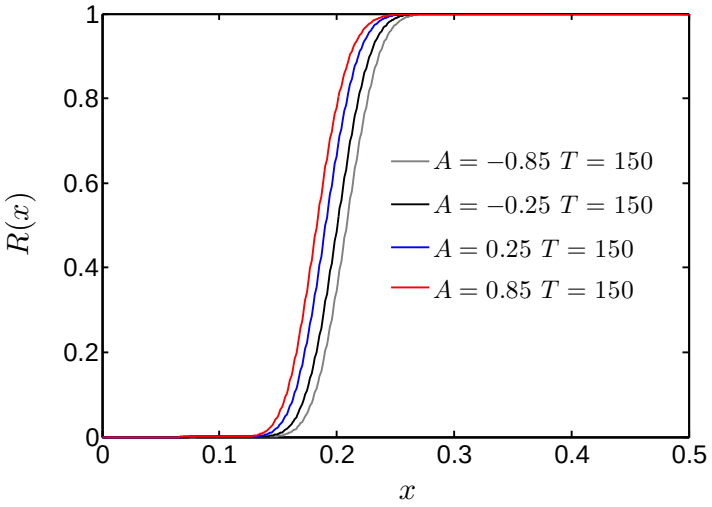


Figure 2.3: Reliability polynomial for GNM networks. Each graph is identified by its assortativity value A and number of triangles T.

2.4 Reliability measures

To elucidate the interplay of network reliability and topological properties, we computed network reliability of the sets of graphs under study. We used an independent edge failure damage model, where the probability for each edge not to fail is equivalent to the probability of disease transmission over that edge. The property used to indicate the reliability of the graph is AR-α with $\alpha = 0.2$, the probability that the cumulative fraction of vertices infected exceeds $\alpha = 0.2$ fraction of the nodes, averaged over all possible initial conditions in which a single vertex is infected. As can be seen in Figure 2.3 the probability of disease reaching to 20% of nodes is smaller for low transmissiblity and it increases with transmissiblity, as expected.

However, to compare different graphs, it is easier to work with scalar values. Thus we further characterized networks' reliability in terms of different features of the reliability polynomial. I contributed in numerical evaluation of this study presented hereafter.

2.4.1 k_{min} and k_{max} minimum and maximum number of edges in reliable subgraphs

As names suggest, k_{min} denotes the minimum number of edges necessary for any subgraph to be reliable. Similarly, k_{max} represents the maximum number of edges for any subgraph to be certainly reliable, i.e. $R(x) = 1$. It is clear that $k_{min} < k_{max}$ since the reliability is monotonically increasing or coherent, i.e. adding an edge to a reliable subgraph does not make it unreliable. Mathematically, the probability P_k is as follows:

$$P_k = \begin{cases} 0 & k \leq k_{min} \\ 0 < P_k < 1 & k_{min} + 1 \leq k \leq k_{max} - 1 \\ 1 & k_{max} \leq k \leq E. \end{cases} \tag{2.4}$$

2.4.2 Average reliability

Average reliability $\langle R(x) \rangle$ is another scalar value based on reliability that is of particular interest in epidemiology context. Average reliability provides the expected probability that at least 20% of the nodes contract the disease averaged over all range of transmission probability, representing the situation that disease transmission rate is unknown. It is easy to see that the average reliability is equivalent to the average probability of selecting a reliable subgraph of any size.

$$
\begin{aligned}
< R(x) > &\equiv \int_{x=0}^{x=1} \sum_{k=0}^{E} R_k x^k (1-x)^{E-k} dx \\
&= \sum_{k=0}^{E} P_k \binom{E}{k} \frac{k!(E-k)!}{(1+E)!} \\
&= (1+E)^{-1} \sum_{k=0}^{E} P_k \\
&\equiv < P_k > \qquad (2.5)
\end{aligned}
$$

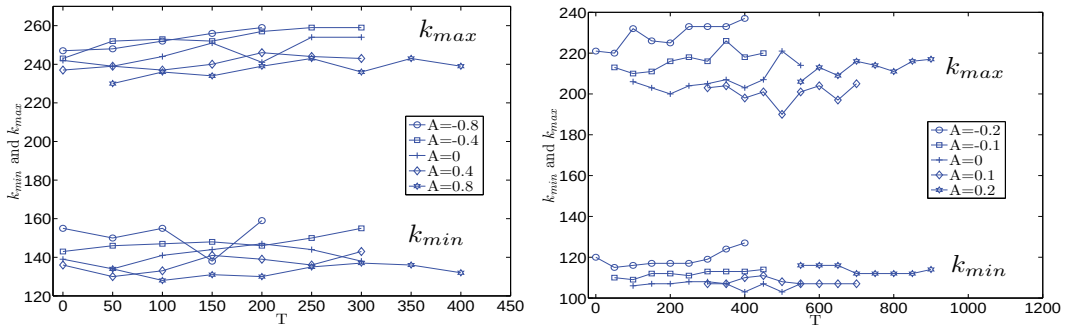
2.4.3 Critical point

Assuming reliability as an order parameter for the network, the first derivative of reliability is the probability that the reliable subgraph starts to percolate if the probability of edge reliability increases from x to $x + dx$. The critical point is where the first order phase transition occurs; i.e. in the thermodynamic limit its first derivative respect to x undergoes a discontinuous change. For finite size graphs however, we define the critical point to be the point where the derivative of the reliability with respect to x attains its maximum value.

2.5 Results

2.5.1 k_{min} and k_{max}

Figure 2.4 shows the minimum and maximum number of edges (k_{min} and k_{max}) needed to obtain reliable subgraphs for GNM graphs. We observe that, in general, both k_{min} and k_{max} decrease as the assortativity increases. Because $k_{min} + 1$ is the minimum number of edges needed to obtain a connected component containing 20% of the vertices, it represents the edge density of reliable subgraphs. Consequently, the edge



(a) GNM graphs

(b) SFL graphs

Figure 2.4: k_{min} (bottom) and k_{max} (top) for disassortative, neutral and assortative GNM and SFL graphs under an AR α reliability rule with α=0.2. A denotes the assortativity of each graph.

density of reliable subgraphs is lower for assortative graphs than for neutral and disassortative graphs. As mentioned in [52], high degree vertices in assortative networks tend to form cliques, which are also called *core groups* in the epidemiological literature. The edge density within the clique is higher than that of the network as a whole. Therefore, a reliable subgraph will first appear with fewer edges within the clique. In disassortative networks, edges tend to connect vertices with dissimilar node degrees. Thus, a reliable subgraph from a disassortative network will first appear with more edges. In other words, reliable subgraphs in assortative networks have lower edge density than reliable subgraphs in disassortative networks as shown in Figure 2.4. We also observe that the number of triangles has more effect on k_{max} than on k_{min}. The number of edges k_{max} increases slightly as the number of triangles increases. Results obtained from SFL graphs are in agreement with results from GNM graphs except when assortativity $A > 0.1$, where both k_{min} and k_{max} increase as the assortativity increases, as shown in Figure 2.4. The edge density of reliable subgraphs is larger for SFL graphs with assortativity $A > 0.1$ than for neutral SFL graphs. In the next subsection, we explain the effect of high assortativity on the reliability of SFL graphs. To understand this phenomenon, note that SFL graphs with near-maximal

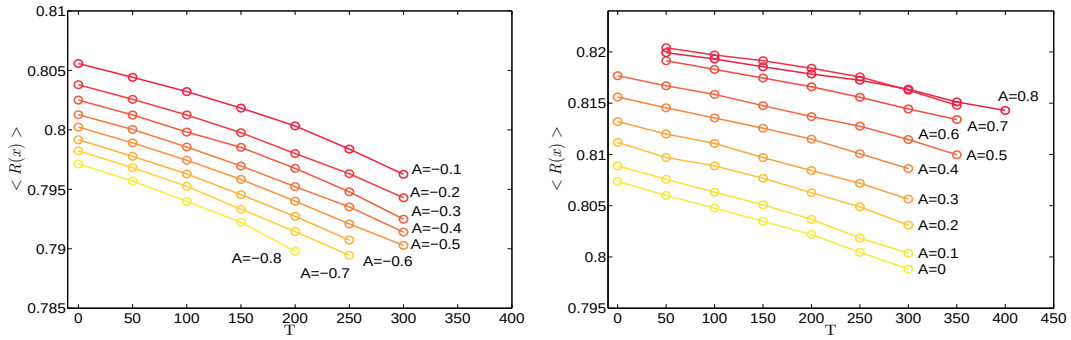
assortativity tend to have large number of triangles, because vertices with similar degrees create cliques. These cliques represent communities with vertices that are strongly connected, while different communities are weakly interconnected. Thus, the number of communities decreases [73] and approaches the number of distinct degree values as assortativity increases for highly assortative SFL graphs. Due to the degree distribution of SFL graphs, the majority of lowest degree vertices belong to a single community. The edge density is lower for this community than for the graph as a whole. Conversely, the communities of high degree vertices contain only a few vertices. Therefore, for reliable subgraphs to appear in communities with high edge density, the reliable subgraphs have to extend across different communities that are weakly interconnected. Consequently, a large number of edges is required to obtain reliable subgraphs from highly assortative SFL graphs. This result causes the critical point to increase with assortativity leading to a decrease in the average reliability.

2.5.2 Average reliability

Figures 2.5(a) and 2.5(b) show the average reliability for GNM networks with negative and neutral or positive assortativity, respectively. We first analyze the influence of assortativity and triangles on the reliability independently.

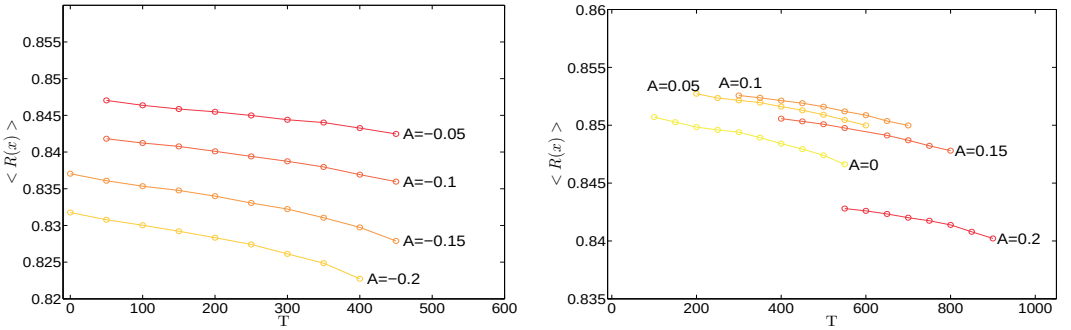
- Effect of triangles on reliability: Network reliability decreases as the number of triangles increases for any assortativity. For the AR-α reliability rule, creating a triangle in an unreliable subgraph by adding a new edge does not make the subgraph reliable since the newly added edge does not increase number of vertices in any connected component. However, if the newly added edge connects a vertex that belongs in one component with a vertex in another component, the probability that the overall subgraph is reliable increases.

- Effect of assortativity on reliability: The more assortative the network is, the



(a) $< R(x) >$ for disassortative GNM graphs $A < 0$

(b) $< R(x) >$ for neutral and assortative GNM graphs $A \geq 0$



(c) $< R(x) >$ for disassortative SFL graphs $A < 0$

(d) $< R(x) >$ for neutral and assortative SFL graphs $A \geq 0$

Figure 2.5: The average reliability $< R(x) >$ and the critical points for disassortative, neutral and assortative GNM graphs under an AR-α reliability rule with $\alpha = 0.2$.

more reliable the network is. We know that reliable subgraphs have lower edge density for assortative graphs than for disassortative graphs i.e. $k_{min}^{assort} < k_{min}^{disassort}$ and $k_{max}^{assort} < k_{max}^{disassort}$. Thus, $\langle P_k^{assort} \rangle > \langle P_k^{disassort} \rangle$. Consequently, using Equation 2.5, $\langle R(x)^{assort} \rangle$ is larger than $\langle R(x)^{disassort} \rangle$.

In contrast to [52], assortative graphs do not always have many cliques. Therefore, we analyze the combined effect of the number of triangles and assortativity on the reliability using six distinct combinations of graph properties:

1) Assortative graphs with few triangles: High degree vertices have high degree neighbors. However, these vertices are not interconnected and hence do not form cliques. Therefore, reliable subgraphs are weakly locally connected. It is hard for a reliable subgraph to percolate among only high-degree vertices because the edge density is lower for the subgraph containing high degree vertices than for the graph as a whole. Therefore, reliable subgraphs expand across not only high degree vertices but also low degree vertices. Due to the assortative property, the majority of vertices will have high degree. Thus, only a small number of edges is required for a reliable subgraph to appear.

2) Assortative graphs with many triangles: The majority of edges are used to create triangles among vertices with similar node degrees. In other words, vertices with similar degrees form weakly interconnected cliques. Reliable subgraphs appear in cliques with high degree vertices due to their large edge density. Because the cliques are highly locally connected, the number of edges in a reliable subgraph is larger for assortative graphs with large number of triangles than for assortative graphs with small number of triangles. In addition, because cliques are only weakly interconnected, it is hard for a reliable subgraph to expand outside the clique.

3) Neutral graphs with few triangles: With equal probability, a randomly selected edge connects vertices with similar degrees or vertices with different degrees. High degree vertices are weakly connected and the subgraph containing them has low

edge density. Being neutral and having few triangles in the graph, a reliable subgraph expands across vertices with a wide range of degrees. Therefore, many edges are required to increase the edge density of a subgraph to become reliable. Thus, a reliable subgraph requires more edges for neutral graphs than for assortative graphs, if they both have few triangles.

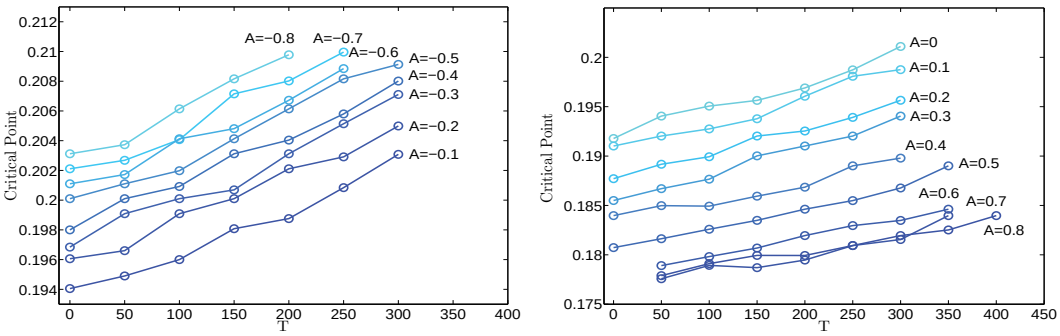
4) Neutral graphs with many triangles: Many triangles exist in the graph without composing cliques. Because the graph is neutral and because triangles do not increase the reliability of graphs, the number of edges needed for a subgraph to be reliable and to expand across the graph is larger for graphs with many triangles than for graphs with few triangles.

5) Disassortative graphs with few triangles: Vertices with different node degrees are connected but do not form cliques. Thus, subgraphs with larger edge density than that of the graph as a whole exist with many edges and vertices. Consequently, in contrast to reliable subgraphs that appear with fewer edges in assortative graphs with few triangles, reliable subgraphs appear with many edges from high density subgraphs.

6) Disassortative graphs with many triangles: Vertices with different node degrees are connected together forming triangles. As discussed above, in finite graphs, triangles do not increase the reliability of graphs.

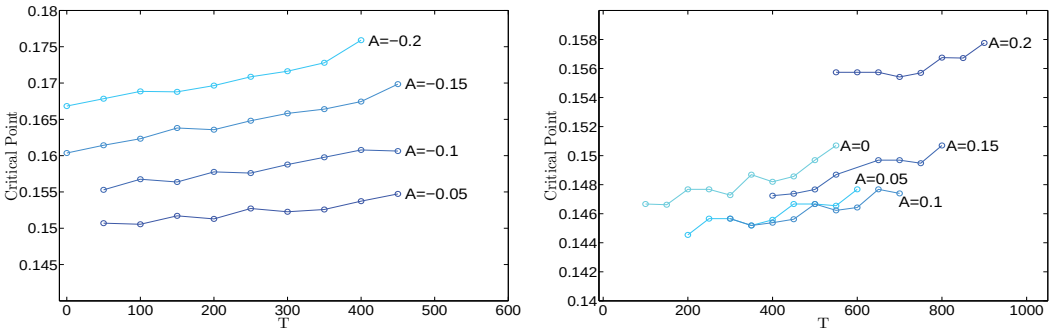
2.5.3 Critical point

The critical point decreases with the increase of degree assortativity. However, the critical point increases with the increase in number of triangles. More edges are required to obtain reliable subgraphs from highly clustered graphs. Also, reliable subgraphs that appear in disassortative networks are more dense than reliable subgraphs from assortative networks. The result is presented in Figure 2.6.



(a) Critical point for disassortative GNM graphs $A < 0$

(b) Critical point for neutral and assortative GNM graphs $A \geq 0$



(c) Critical point for disassortative SFL graphs $A < 0$

(d) Critical point for neutral and assortative SFL graphs $A \geq 0$

Figure 2.6: The average reliability $< R(x) >$ and the critical points for disassortative, neutral and assortative GNM graphs under an AR-α reliability rule with $\alpha = 0.2$.

2.6 Network reliability and scaling

We study the effect of graph size by evaluating the reliability on GNM graphs with fixed average node degrees and sizes V, $2V$ and $4V$. Three different assortativity values are used, while the number of triangles is held constant at 100. The results are summarized in Table 2.1. Let k' be the normalized number of edges with respect to the total number of edges in the graph, e.g. $k'_{min} = \frac{k_{min}}{4E}$ for graphs with $4V$ vertices and $4E$ edges. We observe that the average reliability, k'_{min} and maximum derivative increase as the graph size increases, while k'_{max}, x_c and $k'_{max} - k'_{min}$ decrease as the graph size increases. In addition, results show that the derivative of the reliability with respect to x diverges for larger graph sizes. In other words, the transition from $R(x) = 0$ to $R(x) = 1$ becomes sharper for large graphs than for small graphs. Consequently, at the thermodynamic limit, $k'_{max} - k'_{min}$ converges to 0 i.e. k'_{min} and k'_{max} reach their convergence value k'_{therm}. Thus, $\langle R(x) \rangle$ and x_c converge to $1 - k'_{therm}$ and k'_{therm}, respectively. Therefore, network reliability moves toward a sharp transition for infinite size systems, reflecting a first order phase transition from a region of unreliable subgraphs on one side to a region with only reliable subgraphs on the other side.

2.7 Conclusion

We have argued that the set of coefficients P_k of a reliability polynomial characterize an arbitrary graph's structure in ways that are immediately relevant to specific dynamical processes. We have highlighted various features of reliability that provide useful characterizations of graph structure, e.g. the minimum and maximum number of edges needed to obtain reliable subgraphs, the average reliability and the critical point. We have created a library of graphs with a range of structural properties, i.e.

Table 2.1: Evaluation of average reliability $\langle R(x)\rangle$, k'_{min}, k'_{max}, derivative of reliability at critical point and the critical point x_c for GNM graphs with different graph sizes V, $2V$ and $4V$. Each graph has assortativity A =-0.85, 0 and 0.85 and number of triangles $T = 100$.

$A = -0.85$	V	$2V$	$4V$
$\langle R(x)\rangle$	0.7935	0.7950	0.7972
k'_{min}	0.1522	0.1623	0.1767
k'_{max}	0.2550	0.2440	0.2349
$k'_{max} - k'_{min}$	0.1028	0.0817	0.0582
max $\frac{dR(x)}{dx}$	18.6515	25.4679	38.7695
x_c	0.2066	0.2046	0.2021
$A = 0$	V	$2V$	$4V$
$\langle R(x)\rangle$	0.8049	0.8066	0.8067
k'_{min}	0.1391	0.1563	0.1641
k'_{max}	0.2460	0.2319	0.2228
$k'_{max} - k'_{min}$	0.1069	0.0756	0.0587
max $\frac{dR(x)}{dx}$	18.7144	25.5718	35.2221
x_c	0.1945	0.1925	0.1928
$A = 0.85$	V	$2V$	$4V$
$\langle R(x)\rangle$	0.8173	0.8228	0.8311
k'_{min}	0.1270	0.1462	0.1447
k'_{max}	0.2429	0.2172	0.2016
$k'_{max} - k'_{min}$	0.1159	0.0710	0.0569
max $\frac{dR(x)}{dx}$	18.9097	27.5449	38.7695
x_c	0.1804	0.1759	0.1680

assortativity-by-degree and triangles.

Simulation results for Erdős-Rényi and scale-free-like random graphs in this library reveal that increasing the assortativity and number of triangles has opposite effects on the probability that an epidemic outbreak will achieve an average attack rate of 20%. We found that the required number of edges decreases as the degree assortativity increases; however, the required number of edges increases as the number of triangles increases. In addition, average network reliability increases as the degree assortativity increases but decreases as the number of triangles increases. Moreover, the critical point decreases and the derivative of reliability at critical point diverges as the degree assortativity increases, while the opposite is true for increasing number of triangles. In contrast to assortative GNM graphs, network reliability decreases as assortativity increases for assortative SFL graphs. Furthermore, we have demonstrated that the transition from unreliable subgraphs to reliable subgraphs behaves as expected.

Chapter III

Analyzing Network Reliability Using Structural Motifs

3.1 Abstract

This study uses the reliability polynomial, introduced by Moore and Shannon in 1956, to analyze the effect of network structure on diffusive dynamics such as the spread of infectious disease. We illustrate a representation for the reliability polynomial in terms of what we call *structural motifs*. These are the building block units of the network that play an important role in diffusive processes too. In this study we illustrate how coefficients of reliability polynomial can be interpreted in terms of structural motifs and their overlap. We also demonstrate various structural motifs with different functional importance and how the specific ways in which the structural motifs are interconnected can influence the reliability of the entire network. Therefore, network analysis by means of motif provides a valuable numerical tool to compare different networks' reliability and provide insight into design principles. We illustrate by deriving several general results relating graph structure to dynamical phenomena.

3.2 Introduction

Characterizing networks in a way that is directly relevant to diffusion phenomena on the network is important, but difficult. We argue that the *Network Reliability Polynomial* introduced by Shannon and Moore [48] is a characterization that folds together static measures like degree, modularity, and measures of centrality into precisely the combinations that are most relevant to the dynamics. [45] Conversely, knowledge of reliability can be used to infer structure, in the sense of network tomography. [30] Furthermore, reliability is a useful concept for reasoning more generally about the consequences of structural changes. Colbourn [18] gives a comprehensive introduction to network reliability, while Youssef [79] provides a brief derivation of the form of the reliability polynomial.

A network's *reliability* is the probability that it continues to function after sustaining damage to its component edges and/or vertices. Reliability depends on a parameterized *damage model* $D(x)$, which specifies the probability of damaging a particular set of components, and a parameterized property $\mathcal{P}(\vec{\alpha})$, which specifies what it means for the network to function. Canonical problems in diffusion over fixed networks can be cast in terms of reliability. For example, a damage model $D(x)$ under which edges (respectively, vertices) fail independently with probability $1 - x$ represents bond (respectively, site) percolation. The choice of $\mathcal{P}$ then selects the dynamical phenomenon of interest. For example, the property "contains a connected component including at least a fraction α of the vertices" is appropriate for studying the existence of a giant component. For ease of notation, we often express $\mathcal{P}$ in terms of a corresponding *reliability rule* $r_{\mathcal{P}}$, a binary function indicating whether property $\mathcal{P}$ holds for a given graph. We say that the rule *accepts* a graph g if and only if $r(g) = 1$, that is g has the property $\mathcal{P}$.

Here we introduce four different reliability rules. The first three are the most com-

monly used rules, followed by the last one that could be of use to study percolation problems. To make it easier to understand, we refer to real world problems like designing reliable communication networks and epidemiology which benefit from each of these rules:

1. *Two-Terminal:* a graph is accepted if it contains at least one directed path from a distinguished vertex S (the *source*) to another distinguished vertex T (the *terminus*). Reliability under this rule is the probability that the specified source can send a message to the specified terminus in a damaged network. [39]

2. *K-Terminal:* a graph is accepted if every vertex is in a connected component that includes at least one of K specified vertices. For example, consider a set of K nodes as service centers. Then reliability under this rule is the probability that every operational site in a damaged network is connected to at least one service center. [23]

3. *All-Terminal:* a graph is accepted if it consists of a single connected component. The reliability under this rule, also known as *system reliability* is the probability that every pair of nodes in a damaged network can communicate with each other, or alternatively, that any vertex in a damaged network can broadcast to all the other vertices. [64]

4. *EAR-α:* To understand this rule we discuss the application of bond percolation for the study of the spread of infectious diseases on networks [44]. The probability that an edge does not fail represents the *transmissibility*, i.e., the conditional probability of transmitting infection from one person to another, conditioned on the source being infectious and the destination being susceptible. One of the most important properties of disease dynamics is the *attack rate*, defined as the fraction of the population infected in an outbreak. Most

models of infectious disease exhibit a sharp transition in the attack rate at a critical value of transmissibility. Indeed, this is a percolation phase transition. We can study the relationship between network structure and epidemic outcome by evaluating the reliability under the independent-edge-failure damage model and the parameterized family of properties: "the mean fraction of vertices in a connected component $\geq \alpha$."

Using reliability we can find the relationship between critical transmissibility and all structural information of a network contained in its edge list. In this case the reliability is the probability that the expected attack rate for an outbreak seeded in a single person chosen uniformly at random from the population is at least α. Hence, we call this rule EAR-α.

These rules are all *coherent*. That is, any graph formed by adding an edge to an accepted graph is also accepted.

In a companion to this study [79], we have shown how the concept of network reliability together with an efficient, scalable estimation scheme can shed light on complicated dynamical trade-offs between local structural properties such as assortativity-by-degree and the number of triangles.

Here we introduce a different representation of the reliability polynomial that highlights the role certain network structures play in dynamical phenomena. The representation in [79] is well-suited for computational analysis of networks with up to 10^8 edges, but is analytically tractable only for small networks; the representation presented in the current work is analytically tractable, but computationally feasible only for small networks because of its combinatorial complexity. Thus the results of this paper exactly complement those of the previous paper.

3.3 Reliability polynomials

We use the common notation of $G(V, E)$ for a graph with V vertices and E edges. The graph may be directed or undirected, and it is possible to have multiple edges between two vertices. The vertices and edges may be labeled. The general case of directed edges and labeled vertices and edges is powerful enough to represent extremely complex networks such as interdependent infrastructure networks. Here, without loss of generality, we restrict ourselves to homogeneous networks represented as undirected, unlabeled graphs.

3.3.1 Definition and a common representation

The *reliability* $R(G, \mathcal{P}(\vec{\alpha}), D(\vec{x}))$ of a network G with respect to the property $\mathcal{P}$ under damage model D is the probability that a subgraph of G chosen with probability given by D has property $\mathcal{P}$. A binary rule $r_{\mathcal{P}}$ examines whether subgraph has property $\mathcal{P}$ or not. If it has then:

$$r_{\mathcal{P}(\vec{\alpha})}(g) = 1,$$

otherwise

$$r_{\mathcal{P}(\vec{\alpha})}(g) = 0.$$

i.e. , reliability can be interpreted as the expected value of the reliability rule operator over different subgraphs of G.

We will explicitly include the dependence on the network G and the property $\mathcal{P}$ in notation such as $R(G, \mathcal{P}(\vec{\alpha}), D(\vec{x}))$ only when we wish to distinguish the reliability of two different graphs or two different properties. Moreover, we will not include the damage model itself, but only the values of its parameters $\vec{x}$. Finally, for a homogeneous network in which all edges (or all vertices) fail with the same probability, $\vec{x}$ is a scalar, x. Thus we can write the reliability simply as $R(x)$.

$$R(x) \equiv \sum_{g \subseteq G} r_{\mathcal{P}(\vec{\alpha})}(g) p_{D(\vec{x})}(g) \tag{3.1}$$

For the independent-edge damage model, in which the probability of selecting a subgraph $g \subseteq G$ depends only on the number of its edges, $|g| = k$, and is $x^k(1-x)^{E-k}$, we have:

$$R(x) = \sum_{g \subseteq G} r_{\mathcal{P}(\vec{\alpha})}(g) x^k (1-x)^{E-k} \tag{3.2}$$

We can re-write Equation 3.2 in terms of a sum over subgraphs of different sizes, introduced by Alon et al.[3] as *motifs*:

$$R(x) = \sum_{k=0}^{E} R_k x^k (1-x)^{E-k}. \tag{3.3}$$

R_k is the number of subgraphs of G with exactly k edges that are accepted by the rule. For computational convenience, we often prefer to work with normalized coefficients

$$P_k \equiv R_k / \binom{E}{k}. \tag{3.4}$$

P_k is the *fraction* of subgraphs of G with exactly k edges that are accepted by the rule. $P_k \leq P_{k+1}$ for a coherent rule. P_k can be estimated efficiently via Monte Carlo simulation. [79]

Substituting R_k coefficients in Equation 3.2 with P_ks from 3.4, we can see the resem-

blance to binomial distribution, since $P_k \leq 1$ it is clear that

$$\begin{aligned} R(x) &= \sum_{k=0}^{E} \binom{E}{k} P_k x^k (1-x)^{E-k} \\ &\leq \sum_{k=0}^{E} \binom{E}{k} x^k (1-x)^{E-k} \leq 1 \end{aligned} \tag{3.5}$$

Therefore, $R(x)\colon [0,1] \rightarrow [0,1]$ is a continuous polynomial with only a finite, but possibly large, number of coefficients $R_k, k \in \{0, \ldots, E\}$. That is, the reliability can be thought of as a vector in an $E+1$-dimensional vector space, and the R_k's as the components of the vector in the basis $x^k(1-x)^{E-k}$. There are, of course, many other bases we could choose for this space. An orthogonal basis, such as the first $E+1$ Legendre polynomials, might have useful estimation properties. Here we use another non-orthogonal basis – the functions x^k i.e. the Taylor series expansion – because of its simplicity and its attractive interpretation. There is a unique mapping from coefficients in one basis to those in the other, which can be derived by expanding the factor $(1-x)^{E-k}$ in Equation 3.3:

$$\begin{aligned} R(x) &= \sum_{k=0}^{E} R_k x^k \sum_{m=0}^{E-k} \binom{E-k}{m} (-1)^m x^m \\ &= \sum_{k=0}^{E} R_k \sum_{l=k}^{E} \binom{E-k}{l-k} (-1)^{l-k} x^l \\ &= \sum_{l=0}^{E} (-1)^l x^l \sum_{k=0}^{l} (-1)^k R_k \binom{E-k}{l-k} \\ &= \sum_{l=0}^{E} N_l x^l \end{aligned} \tag{3.6}$$

where

$$N_l \equiv (-1)^l \sum_{k=0}^{l} (-1)^k \binom{E-k}{l-k} R_k. \tag{3.7}$$

The N_l coefficients are signed integers. In section 3.3.3 we will explain how we can

interpret these coefficients.

3.3.2 Structural motifs

We can express the reliability polynomial in terms of overlaps among certain distinguished subgraphs. These subgraphs are the $\mathcal{P}$-minimal subgraphs of G. A graph g is $\mathcal{P}$-minimal if and only if:

1. g has property $\mathcal{P}$; and
2. there is no proper subgraph $g' \subset g$ that has property $\mathcal{P}$.

Obviously, whether a graph is $\mathcal{P}$-minimal or not depends on the property $\mathcal{P}$, or equivalently here, the reliability rule. For example, $\mathcal{P}$-minimal graphs under a Two-Terminal rule are paths from S to T with no extraneous edges, i.e., no loops or dead ends; under the All-Terminal rule, they are spanning trees. In general, the rule selects a distinctive topological pattern (e.g. path, spanning tree) that may occur many times in a given graph, i.e., if we consider all subgraphs of a distinctive pattern to be the *motif*s introduced by Alon et al.[3] then reliability rule selects a subset of these motifs that has the property $\mathcal{P}$. We refer to subsets generated by a particular rule as *structural* motifs because, as we will demonstrate, they are the structural elements of the network that completely determine the occurrence of dynamical phenomena of interest, as specified by $\mathcal{P}$. One advantage of using this representation is that the contribution of the structural motifs to the reliability is known exactly for all $\mathcal{P}$ and all D, as is shown below.

To describe this, we apply two terminal reliability rule with *source* and *terminus* nodes as its parameters on a small toy network, depicted in Figure 3.1. For this rule structural motifs are simple paths (including but not limited to shortest paths)

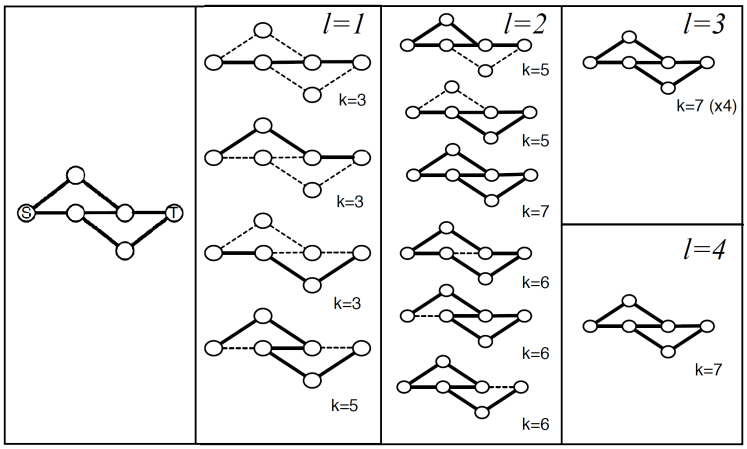


Figure 3.1: Example calculation of two-terminal reliability in the toy network in the left panel. The second panel from the left shows the four motifs; the third shows all unions of two motifs. There are $\binom{4}{3} = 4$ unions of 3 motifs, all of which give the entire graph as shown on the top of the right panel; there is one union of all motifs, which is also the entire graph. The number next to each union of motifs gives its size.

connecting these two nodes. The structural motifs and their overlap can be seen in Figure 3.1.

In another example, we consider the two-terminal reliability rule on a two dimensional 4x4 grid. For a given source node we look at two different terminus nodes to illustrate the dependence of structural motifs on parameters of the reliability rule. Several structural motifs of different sizes for each set of selected sources and termini are shown in Figure 3.2.

3.3.3 A physical interpretation for N_k

An Inclusion-Exclusion argument lets us express the reliability polynomial in terms of overlaps among structural motifs. We define a *structural motif* to be a minimal subgraph accepted by the rule, i.e. a minimal set of edges that guarantees that the graph containing them is accepted. Formally, a structural motif is a set $\mathcal{M} \subseteq \mathcal{E}$ of edges such that the subgraph $g(V, \mathcal{M})$ is accepted, i.e. $r(g) = 1$, and the subgraph

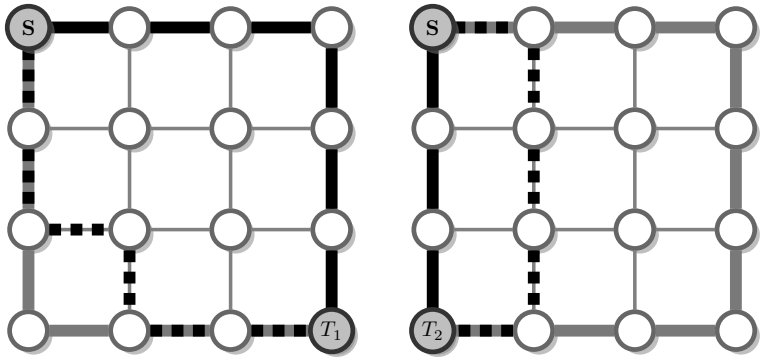


Figure 3.2: Structural motifs for the two-terminal reliability rule are shown for same source node S and two different target nodes T_1 and T_2. We can see that structural motifs of the same network vary with the parameters of property $\mathcal{P}$. In the left grid three structural motifs of sizes 6 are shown: note that black dashed path and gray line have all but two edges in common. For the grid on the right, black normal, black dashed and gray paths represent motifs of sizes 3, 5, 9 respectively. The last two have two edges in common.

$g'(V, \mathcal{M}')$ is *not* accepted if $\mathcal{M}' \subset \mathcal{M}$ (minimality).

Let us consider several important examples. In general, rules that place constraints on the size of connected components admit trees as structural motifs. For example, a spanning tree is a minimal acceptable subgraph, and thus a structural motif, for all-terminal reliability. Note that the use of "minimal" means that each edge in the subgraph is required for reliability, and *not* that the subgraph itself contains no more edges than any other reliable subgraph. For example, for 2-terminal reliability, any loop-free path – not just the shortest paths – from vertex S to vertex T is a structural motif. For a coherent rule, adding an edge to a path creates an acceptable, but no longer minimal, graph.

We motivate the Inclusion-Exclusion argument with a few straightforward examples.

3.3.3.1 Example 1: A single structural motif

Suppose the graph contains only one structural motif and that it is a set of k_0 edges. For example, for 2-terminal reliability, suppose that there is exactly one path between S and T, and that it has length k_0. Then the motif will occur exactly once among all subgraphs of size k_0. For $k > k_0$, think of "using up" k_0 edges on the structural motif. This leaves $E - k_0$ other edges from which, because the rule is assumed to be coherent, any set of $k - k_0$ produces an acceptable subgraph of size k. Hence for this case

$$R_k = \begin{cases} 0 & k < k_0 \\ \binom{E-k_0}{k-k_0} & k \geq k_0 \end{cases} \tag{3.8}$$

3.3.3.2 Example 2: Two disjoint structural motifs

Suppose the graph has exactly two structural motifs, and that both have k_0 edges, and no edge is in both. Arguing as above, for $k < 2k_0$, R_k is simply twice what it is for the case of a single motif. But when $k = 2k_0$, the subgraph that consists of the union of the two motifs will have been counted twice instead of once. Similarly, for $k > 2k_0$, the number of graphs overcounted is given by assigning $2k_0$ of the edges and choosing the remaining $k - 2k_0$ in the subgraph from among the remaining $E - 2k_0$ in the base graph. Hence:

$$R_k = \begin{cases} 0 & k < k_0 \\ 2\binom{E-k_0}{k-k_0} & k_0 \leq k < 2k_0 \\ 2\binom{E-k_0}{k-k_0} - \binom{E-2k_0}{k-2k_0} & 2k_0 \leq k \end{cases} \tag{3.9}$$

3.3.3.3 Example 3: Three disjoint structural motifs

Suppose the graph has exactly three structural motifs, that all three have k_0 edges, and that the three edge sets are disjoint. Again, when $k_0 \leq k < 2k_0$, each motif generates $\binom{E-k_0}{k-k_0}$ different reliable subgraphs, and for $2k_0 = k$, three of these subgraphs are counted twice. But in this case, when k reaches $3k_0$, the subgraph consisting of all three motifs is first included three times (once for each motif), then excluded three times (once for each pair of motifs) with the net result that it must be included again:

$$R_k = \begin{cases} 0 & k < k_0 \\ 3\binom{E-k_0}{k-k_0} & k_0 \leq k < 2k_0 \\ 3\binom{E-k_0}{k-k_0} - 3\binom{E-2k_0}{k-2k_0} & 2k_0 \leq k < 3k_0 \\ 3\binom{E-k_0}{k-k_0} - 3\binom{E-2k_0}{k-2k_0} + \binom{E-3k_0}{k-3k_0} & 3k_0 \leq k \end{cases} \tag{3.10}$$

3.3.3.4 Example 4: N disjoint structural motifs

Suppose the graph has exactly N structural motifs, that all have k_0 edges, and that the N edge sets are disjoint. (Thus $N \leq E/k_0$.) Applying the arguments above, with the convention that $\binom{a}{b} = 0 \ \forall b < 0$, gives:

$$R_k = \sum_{i=1}^{N} (-1)^{i+1} \binom{N}{i} \binom{E - ik_0}{k - ik_0} \tag{3.11}$$

3.3.3.5 Example 5: Two overlapping structural motifs

Suppose the graph has exactly two structural motifs, that both have k_0 edges, and that the number of edges in the union of the two is $k_0 + \Delta$. Arguing as in Example 1,

we get a similar result, with $2k_0$ replaced by $k_0 + \Delta$:

$$R_k = \begin{cases} 0 & k < k_0 \\ 2\binom{E-k_0}{k-k_0} & k_0 \leq k < k_0 + \Delta \\ 2\binom{E-k_0}{k-k_0} - \binom{E-k_0-\Delta}{k-k_0-\Delta} & k_0 + \Delta \leq k \end{cases} \tag{3.12}$$

3.3.3.6 The general case

Suppose the graph has exactly N structural motifs. As above, its reliability polynomial will be determined by the size of each structural motif and the overlaps among them. Define $\tilde{N_k}^{(l)}$ as the number of combinations of l structural motifs whose union contains exactly k edges. Also, define

$$\tilde{N_k} \equiv \sum_{l=1}^{N} (-1)^{l+1} \tilde{N_k}^{(l)} \tag{3.13}$$

Then arguing as above gives

$$R_k = \sum_{k'=0}^{k} \tilde{N_{k'}} \binom{E-k'}{k-k'} \tag{3.14}$$

3.3.4 Reliability in terms of structural motifs

Given the rather complicated relationship between R_k and $\tilde{N_k}$ in Equation 3.14, it is surprising that $R(x)$ can be expressed very simply in terms of $\tilde{N_k}$. Consider the contribution of a single structural motif of size k_0 to $R(x)$. Using Equation 3.14,

Table 3.1: By inspection, we have the values for $N_k^{(l)}$ and the values of N_k and R_k as given by Equations 3.13 and 3.14.

l	k	$N_k^{(l)}$	N_k	R_k
1	3	3	3	3
	4		0	12
1	5	1		
2	5	2	-1	17
2	6	3	-3	7
2	7	1		
3	7	4		
4	7	1	2	1

$R_k = \binom{E-k_0}{k-k_0}$. This set of coefficients determines $R(x)$

$$\begin{aligned} R(x) &\equiv \sum_{k=0}^{E} R_k x^k (1-x)^{E-k} \\ &= \sum_{k=0}^{E} \binom{E-k_0}{k-k_0} x^k (1-x)^{E-k} \\ &= x^{k_0} \sum_{k=k_0}^{E} \binom{E-k_0}{k-k_0} x^{k-k_0} (1-x)^{E-k_0-(k-k_0)} \\ &= x^{k_0} \sum_{k'=0}^{E-k_0} \binom{E-k_0}{k'} x^{k'} (1-x)^{E-k_0-k'} \\ &= x^{k_0} \end{aligned} \tag{3.15}$$

Since the effect of each structural motif, and each motif overlap, is additive on R_k, we can reduce the general case to sums like the above, so we immediately find:

$$R(x) = \sum_{k=0}^{E} \tilde{N}_k x^k \tag{3.16}$$

Thus the $\tilde{N}_k$ defined in Equation 3.13 are indeed the same coefficients as those introduced in Equation 3.7.

Table 3.2: Number of structural motifs and their overlap for the two terminal reliability rule on the left grid in Figure 3.2.

k	N_k^1	N_k^2	N_k^3	N_k	R_k
5	0	0	0	0	0
6	20	0	0	20	20
7	0	0	0	0	360
8	36	30	0	6	3066
9	0	84	0	-84	16332
10	48	146	144	10	60670

3.3.5 Alternative damage models

The reasoning above is all done in the context of the usual edge damage model introduced by Moore and Shannon. This damage model is appropriate for studying bond percolation. An entirely analogous set of arguments applies to a vertex damage model, in which a set of k *vertices* is chosen uniformly at random, producing a unique subgraph containing all the edges whose endpoints are both in the selected set of vertices. This damage model is appropriate for studying site percolation. Coefficients analogous to P_k and N_k can be derived (substituting the number of vertices V for the number of edges E wherever it appears) and structural motifs can be defined in terms of vertex removal instead of edge removal. The physical interpretation of N_k in terms of these structural motifs is the same. It is likely that there are many other damage models with these properties. Here, we consider only the edge damage model, because it serves to illustrate the role of structural motifs and its analysis is simpler.

3.4 Structural motifs to find edge importance

In section 3.3.2 we demonstrated how structural motifs for the 2 dimensional grid depend on parameters of the two terminal reliability rule i.e. S and T. Here we explain how this fact can effect edge importance based on reliability rules. We computed reliability for two-terminal reliability rule for two cases on two dimensional grids in Figure 3.2. It is clear that grid is more reliable for the reliability rule parameters in the right, i.e., it is more probable to have a path from S to T_2 than to T_1. This can be seen in Figure 3.3. Next we remove two out of three edges on the shortest path connecting S and T_2 and we compute the reliability for both cases again. We see that reliability of the grid decreases more for the left case as expected. This result suggests to employ structural motifs for finding most important edges in a way that reflects the choices of parameters for the reliability rule.

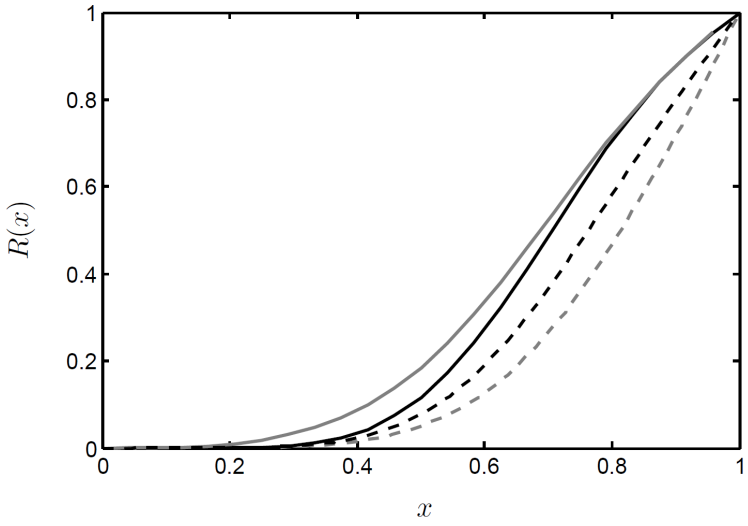


Figure 3.3: Reliability polynomial for two terminal reliability black curve represents reliability for $S-T_1$, gray curve for $S-T_2$ and dashed curves represent the reliability polynomial after removing two edges.

The reliability polynomials of different graphs may intersect multiple times. [13] This means that, for different values of the damage model's parameters $\vec{x}$, the relative reliability of the two graphs switches signs. If both graphs are subgraphs of the

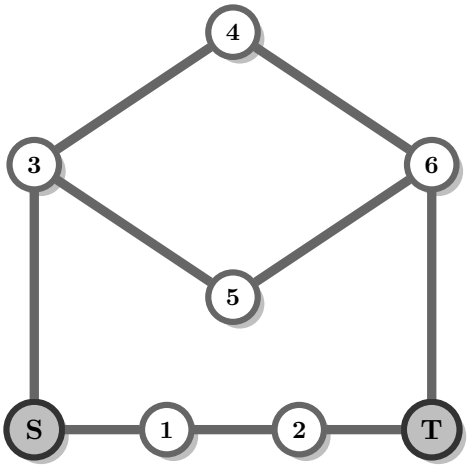


Figure 3.4: Toy graph illustrating the possible x-dependence of an edge's importance.

network of interest with the same number of elements removed, then clearly the ranking of the elements is different for different parameter values. This is perhaps surprising, but it is an important feature of this approach compared to, for example, a graph statistic that is independent of $\vec{x}$.

For example, consider the graph in Figure 3.4 under $S-T$ reliability for the indicated S and T. First, we write down the reliability for the graph by inspection from its structural motifs. There are three motifs, $A \equiv S12T$, $B \equiv S356T$, and $C \equiv S346T$. The first has size 3; the second and third, size 4. The second and third overlap in two edges, but are disjoint from the first. There are no edges that do not appear in any structural motif. Taken together, this gives:

$$R(x) = x^3 + 2x^4 - x^6 - 2x^7 + x^9. \tag{3.17}$$

(Note that $R(x)$ satisfies the constraints $\sum N_k = 1$ and $\sum |N_k| = 2^m - 1$, where $m = 3$ is the number of structural motifs.) By symmetry, we expect the three edges $S1$, 12, and $2T$ to be equally important, and also the pair $S3$ and $4T$, and finally the four edges 35, 36, 54, and 64. Which edge is most important? A moment's thought shows that any edge from the last four is less important than any other edge. The

real choice is between $S1$, which is part of a single structural motif of size 3, and $S3$, which is part of *two* structural motifs of size 4. We consider the reliability $R_1(x)$ of the graph after removing edge $S1$, leaving motifs B and C, and the reliability $R_2(x)$ after removing edge $S3$, leaving only motif A. Again by inspection, these are:

$$R_1(x) = 2x^4 - x^6 \tag{3.18}$$

$$R_2(x) = x^3. \tag{3.19}$$

By definition, the importance of the edges is $I_{S1}(x) \equiv R(x) - R_1(x)$ and $I_{S2}(x) \equiv R(x) - R_2(x)$. Hence, the rank of the edges switches if the polynomial $I_{S1}(x) - I_{S2}(x) = R_2(x) - R_1(x)$ changes sign. In fact, this polynomial has a zero in the interval $[0, 1]$. That is,

$$R_2(x) - R_1(x) = x^3(1 - 2x + x^3) \begin{cases} > 0 & \text{for } x < 0.618... \\ < 0 & \text{for } 0.618... < x \end{cases} \tag{3.20}$$

3.5 Applying reliability concepts to other network analysis problems

The representation of the reliability polynomial in terms of structural motifs provides a convenient organizing principle for thinking about general network analysis problems. As one example, consider the tradeoffs between two systems: one with only a few completely redundant reliable subsytems and another with more, but only partially redundant, ones. To study this we consider two extreme cases of overlap. One contains r_1 structural motifs of size k_1, any two of which differ by only two edges. They are thus built using a total of $2r_1 + k_1 - 2$ edges. The reliability of this

combination can be written as:

$$R_1(x) = \sum_{i=1}^{r_1} (-1)^{i+1} \binom{r_1}{i} x^{k_1 + r_1(i-1)}. \tag{3.21}$$

Using the same number of edges we can construct $r_2 = \frac{2r_1+k_1-2}{k_2}$ motifs of size k_2 that are completely disjoint. The reliability of this combination of motifs is:

$$R_2(x) = \sum_{i=1}^{r_2} (-1)^{i+1} \binom{r_2}{i} x^{ik_2}. \tag{3.22}$$

Knowing the reliability for these two cases, we are able to compare the reliability of networks with different configurations of structural motifs of different sizes. As an example we compared the reliability of a network composed of 20 motifs with 18 edges that are different from one another only in two edges with a network of 4 completely disjoint motifs of size 6. Figure 3.5 shows the reliability curves for these two networks and their difference as a function of x. The analysis shows that the network of disjoint motifs is more reliable for smaller values of x while the opposite is true for larger x values.

This approach could also be used to estimate the number of spanning trees in a graph. A spanning tree is a subgraph of the network that includes all vertices [16, 21, 14]; the number of spanning trees can be estimated by evaluating the All-Terminal reliability. Another problem that can be addressed using this method is to identify chordless loops of various sizes in a network. A chordless loop is a sequence of vertices with more than three vertices if for all $i = 1, \cdots, k$ there is exactly one link from vertex v_i to v_{i+1} and there is no other link between any two of the vertices in this sequence [69]. Recent studies on ecological networks have discovered the existence of many chordless cycles in these networks [37], therefore enumeration of all chordless cycles can make a significant impact on understanding the structure of these networks. An

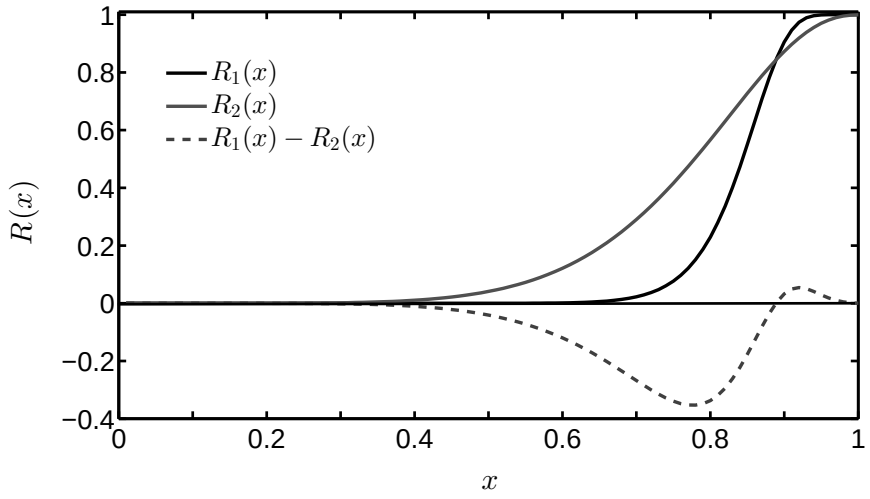


Figure 3.5: Comparing the reliability of a network with many overlapping structural motifs with that of a network with a few disjoint motifs.

appropriately-designed reliability rule can be used to count the number of chordless cycles of different sizes.

3.6 Conclusion and future work

In this study we focused on the representation of the reliability polynomial in terms of structural motifs. We have shown that network reliability is simply related to the number of edges in unions of structural motifs N_k (3.16). Whereas the coefficients P_k of $x^k(1-x)^{E-k}$ are easy to estimate and hard to work with analytically, the coefficients N_k of x^k are hard to estimate but easy to work with analytically. To demonstrate this, we have derived closed-form expressions for N_k for several types of graphs. The resulting expressions were confirmed by numerical estimation. We anticipate that this approach can lead us to a measure of edge centrality that relates the importance of an edge to the frequency of its appearance in different structural motifs [26]. While we can use numerical simulation to study specific large, realistic networks – including

epidemiology on social networks [25, 59, 22] – we can use the notion of structural motifs to understand the differences between networks that are discovered in simulation. We expect this approach to be particularly useful in studying the stability and robustness of interconnected networks [31, 15].

Chapter IV

Reliability-based criticality measure

Portions of this chapter were originaly published in the Proceedings of the second workshop on complex networks and their applications, SITIS 2013, pages 641-648. Reprinted, with permission from [26] with minor changes.

4.1 abstract

Contact network analysis has been widely used in the study of epidemiology. One pioneering study in this area emerged at the onset of HIV, when epidemic information revealed that contact with an infectious individual could cause disease transmission. The Auerbach et. al. [4] study uncovered the importance of sexual activity as a factor in the spread of AIDS and ascribed an increased risk corresponding to an increase in the number of contacts. Key observations such as this one were not feasible without access to contact-based studies of empirical data obtained from face-to-face interviews with AIDS patients. Over the ensuing years several studies were conducted to investigate the interplay between contact patterns and HIV transmission and to design intervention strategies that could assist in controlling the epidemic. One main avenue of research was developed to identify the individuals that play the most impor-

tant roles in spreading the infection. Numerous measures from network theory have been applied to determine a given individual's prominence in a network. [65] However, most of these measures suffer from an important limitation: they are merely statistical measures of network connectivities which regardless of the dynamical process taking place on the network, assign the same ranking of nodes. In this study we propose a measure to mitigate this shortcoming by incorporating the dynamical phenomena of interest—in this case disease propagation— into the evaluation. Our measure relies on network reliability, a theoretical framework to study diffusive phenomena on networks, which can consequently be used to quantify disease spread on these networks. We use this methodology to assess the most critical entities (nodes/edges) in a network and further explain its applicability to epidemiology. We used data from a well-known study of a population presumed to be at high risk of HIV, called the *Colorado Spring Project 90.* [61] We illustrate the effectiveness of our centrality measure through computation and comparison with other centrality measures of the network obtained from this data set.

4.2 Introduction

4.2.1 Important problems in infectious disease epidemiology

The spread of infectious disease through a population is well-represented as a diffusion process on a contact network whose vertices represent hosts and whose edges represent opportunities for transmission. [25] For example, the network version of the commonly-used $Susceptible - Infectious - Recovered$, or $S - I - R$, model is equivalent to bond percolation. [33] We can view network structure as a set of model parameters alongside the usual biological parameters of the host-pathogen interaction such as incubation period or susceptibility. This perspective raises the possibility of

applying new tools to infectious disease epidemiology. Two questions in particular can be viewed as specializations of network characterization and design problems:

- **Q1 Characterization:** How can we parameterize aspects of network structure that have important dynamical consequences?
- **Q2 Design:** What structural changes – i.e. what changes to these structural parameters – most cost-effectively control the consequences of an outbreak?

These questions have, of course, been raised many times in many domains besides epidemiology. Proposed parameterizations include most graph statistics, such as distributions of degree, clustering, shortest-path lengths, and eigenvalue spectra of the adjacency and related matrices. Note, however, that answers to these questions depend on a cost function that defines the "consequences", whereas graph statistics do not. We develop and illustrate a theoretical framework and practical methodology for answering these questions when the costs are properties of dynamical processes on the network. We propose using Moore and Shannon's concept of network reliability [48] to characterize the network, and provide a general algorithm for finding a set of edge or vertex deletions that achieves maximal change in reliability.

For model results to inform policy, the limitations of the models must be well-understood. In the case of networked epidemiology models, that requires understanding the sensitivity of the model to uncertainty in network specification. Moreover, since the biological parameters are largely beyond our control, the only "control surface" for epidemiology is the structure of the contact network. For example, isolating or vaccinating an individual with a perfect transmission-blocking vaccine removes the corresponding vertex from the network; closing a location where people come into contact, such as a school, removes edges corresponding to all the contacts that would have occurred there. Answers to the characterization and design questions will permit a better understanding of sensitivity and control in network epidemiology.

The human immunodeficiency virus (HIV), which leads to immunodeficiency syndrome (AIDS), emerged in 1981 and has continued to be a major cause of suffering and mortality in several parts of the world. HIV is a primarily sexually-transmitted disease which can also be transmitted by contaminated needles and blood transfusions. These methods of movement through the population–sexual contact and the sharing of needles by users of injectable drugs–suggest that contact patterns are important factors in the spread of HIV. Consequently, a network representing the contact pattern has been integrated in the modeling of this disease. In this report, we use a network built from a HIV survey, to answer **Q1**, more particularly, to gauge the central nodes in the network. Central nodes are those which play the most important role in the network for a given question. For example in epidemiological studies they are the nodes which play dominant roles in spreading the contagious disease across populations. In other disciplines, central nodes may be important in different contexts, for instance in the robustness of transportation or telecommunication networks, where a disruption has a greater impact when central nodes are involved. Here we are interested in epidemiological implications of node centrality; our goal is to find the nodes whose removal inhibits the spread of the infectious disease across the population.

4.3 Method

4.3.1 Network epidemiology

Various mathematical models have been used to study the spread of infectious diseases in populations. The most commonly used are compartmental models, where each individual is in one of the different stages of the disease. For instance in a three compartment model known as **SIR**, each person could either be susceptible, infected

or recovered. For a comprehensive review of epidemic modeling, we refer the reader to an article by H. T. Hethcote. [36] These models, however, are based on mean field assumptions that consider a homogeneously mixed population. For instance, each infected person has an equal chance of contacting anyone else in the population and every individual is exposed to the same number of infected individuals and thus have the same chance of becoming infected.

This means that, compartmental models fail to capture the true nature of disease propagation in a heterogeneous population. To account for heterogeneity, mathematical modeling efforts have incorporated contact networks into the evaluation of disease transmission. This approach, based on empirical studies, locates individuals in more realistic social settings and provides a better description of the contact pattern based on observed relationships. In a study of the spread of measles in a high school [24], Rapoport and Yuan compared a network constructed from the actual population with a random contact model. They showed that, in the actual network, epidemic spreads more slowly and reaches fewer people, emphasizing how social ties can act as a roadblock to the spread of infection. Moreover, using contact patterns helps to determine who is at higher risk of contracting an infection or which individuals can transmit the disease to more people in the population. Contact networks provide the information necessary to design intervention techniques in a population and can help determine the efficacy of prevention programs.

The spread of infectious disease through a population is represented as a diffusion process on a contact network. Each individual from the population is represented as a node which, if infected, can infect its neighboring nodes. Thus edges represent opportunities for disease transmission [25]. From here two slightly different approaches can be considered. First, it is intuitively known that not every contact with infected individuals leads to disease transmission. This can be modeled as a probabilistic process, where links connecting nodes in the network are only transmitting disease

with some probability p. This is similar to the **bond percolation model.** Secondly we can use the **site percolation** to model the fact that individual may contract the disease with some probability.

Therefore, the network version of the commonly-used $Susceptible - Infectious - Recovered$, or SIR, model is equivalent to bond percolation. [33]

In this study we investigate the spread of HIV on a network built from survey data. Data for this study was gathered as a part of a Centers for Disease Control and Prevention funded study focused on controlling sexually transmitted diseases.

4.3.2 Network model

The Colorado Springs Project 90 was implemented from 1988 to 1992 in Colorado Springs situated to the south of Denver, Colorado. The focus of the study was on the population presumed to be at high risk for HIV transmission: prostitutes, their partners (paying or nonpaying), injecting drugs users (IDUs) and their sexual partners. A sample of 595 individuals at high risk was interviewed annually over the course of the study, and information gathered on respondents' demographic information, alongside with their social, sexual, drug use and needle sharing contacts. Respondent could report both on their immediate contacts as well as the relationship their contacts have with others. Study details, including the sampling design, have been reported by Klovdahl et. al. [43]. Data from this study has been used extensively to investigate the impact of contact patterns on HIV transmission.

For the purpose of the current study, we focus only on the first interview conducted in 1988, to avoid the impact of dynamical changes in the contact pattern. We have also constrained the network to contain only the direct contacts reported by respondents. As a result we have a network with $V = 3827$ vertices representing respondents and $E = 4027$ edges representing the contacts between them. However, as can be seen

in Figure 4.1 the network is composed of disconnected components. Since no disease transmission is possible between these disconnected component, we restricted the network to include the giant connected component only, with $V = 2274$ vertices and $E = 2667$ edges.

The budget for controlling an outbreak of infectious disease is rarely large enough to reduce the probability of disease spread to 0. Instead, we try to prioritize changes to the network structure, for example ranking edges or nodes to target for removal. In order to validate these measures, we need a quantitative observable that can capture how different nodes may differentially influence the spread of the epidemic. To obtain this we pose network reliability as an observable that quantifies the diffusive processes on networks. We illustrate that network reliability can directly capture the impact of network structure on dynamical processes diffusing on a network and therefore provide a powerful tool to characterize the structure of graphs. Here we use the reliability rule "AR-α" which accepts graph g if and only if the mean square size of connected components is $g \geq \alpha V$; This rule has an important interpretation as the expected probability that an outbreak seeded in a single vertex chosen uniformly at random will eventually infect 20% or more of the vertices. We would like to design social contact networks that transmit disease as unreliably as possible. We evaluated reliability of the Colorado Spring network under this rule for $\alpha = 0.1$ and $\alpha = 0.2$. As can be seen in Figure 4.2, the network is less reliable for $\alpha = 0.2$, i.e. it is less probable to obtain a graph that contains more than $0.2V$ of the nodes than to get one with $0.1V$ nodes. Furthermore, we have introduced an interpretation for the reliability polynomial's coefficients in terms of minimal subgraphs which we refer to as *structural* motifs [42, 26]. Structural motifs support analytical reasoning about the consequences of changes in network structure.

Ranking edges by betweenness and removing the top-ranked ones is often suggested as a heuristic solution to the network design problem. Network reliability can be used

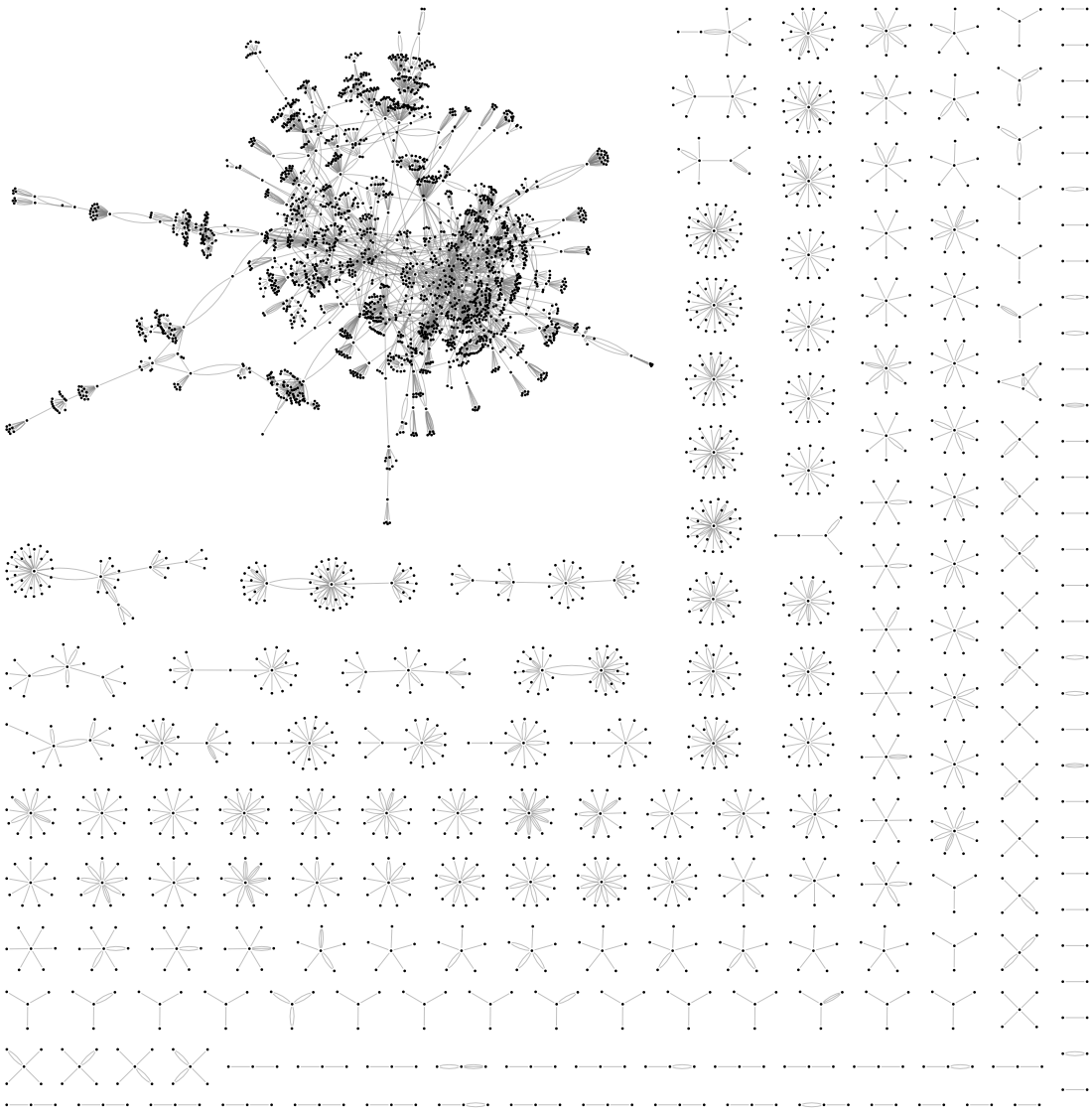


Figure 4.1: Network structure for Colorado Spring Project 90, including respondents' immediate contacts reported during the first round of interviews.

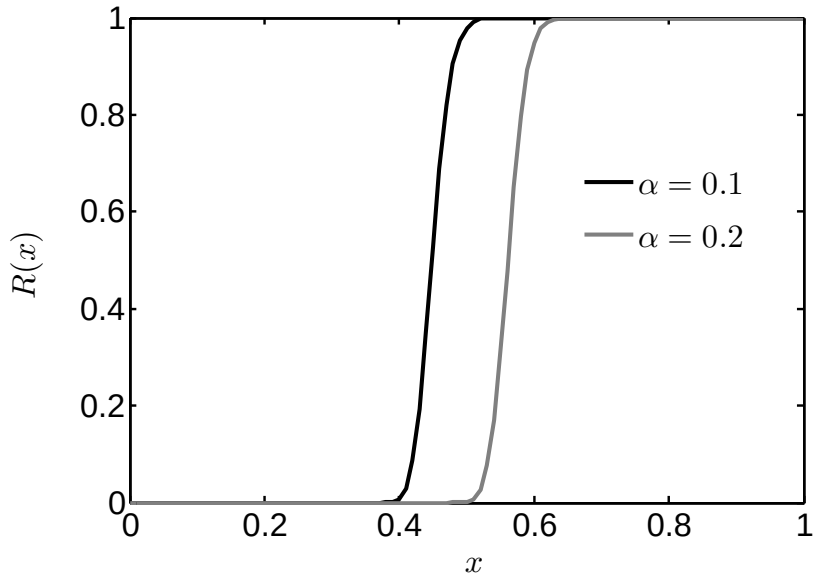


Figure 4.2: Network reliability for Colorado Spring Project 90 under reliability rule "AR-α" for two values of α, 0.1 and 0.2. Network is less reliable for $\alpha = 0.2$, in other words for each given x it is less probable to obtain a subgraph containing 0.2 fraction of the nodes than to get one with 0.1 fraction of nodes.

to generalize the notion of betweenness to include the specific dynamical phenomena and cost functions of interest. Representing network reliability in terms of structural motifs, we demonstrate that the generalized betweenness can in principle be used to solve the network design problem. We evaluate the feasibility of approximating this solution on large networks using our distributed reliability estimation tool.

4.4 A reliability-based measure of centrality

We are interested in a node's "importance" in the sense of criticality here.[58] The notion of a node being critical is perhaps most familiar for ST-reliability, where we even say colloquially that a vertex or edge is on the *critical path* if it must be traversed to get from S to T, i.e. if its removal ensures there is no path from S to T. If there are multiple paths from S to T, no single path is critical, and the binary notion of

criticality must be extended to reflect the fraction of paths that contain the node. The notion of criticality must be further extended to account for the fact that different paths contribute differently to the dynamics. For example, if there are two paths of different length between S and T, the shorter path is often more important than the longer one. Betweenness is the concept which captures these extensions. We extend the intuitive notion of being on a critical path to a notion of criticality for arbitrary properties of dynamics. See Page and Perry for a related approach. [58]

In our more general context, structural motifs play the role of critical paths. We correctly account for the possibility of multiple structural motifs with different contributions to the dynamics with this definition: the *criticality* of a set of nodes is the difference in reliability between the graph G in which they appear and the graph $G^* \subset G$ in which they do not.[1] Since reliability depends on the rule, the damage model, and its parameter values $\vec{x}$, criticality also depends on these things. The reliability polynomial weights each structural motif in precisely the manner it contributes to the dynamical phenomena of interest, as specified by the reliability rule, in the parameter range of interest, as specified by $\vec{x}$. The notions above can be written more formally.

4.4.1 Definition

Let $R^*(G, x, v)$ be the reliability of the graph $G - v$, i.e. G with node v removed: $R^*(G, x, v) = R(G - v, x)$. R^* can be represented in polynomial form as in Equations 3.3 and 3.13. We denote the coefficients appearing in these representations of

[1] A similar definition for edge criticality can be made, with obvious changes in the following remarks.

R^* by $P^*(v)$, $N^*(v)$, etc. Then we define the criticality of node v as

$$\begin{aligned}\Delta_v(x) &\equiv R(G,x) - R^*(G,x,v) && (4.1)\\ &= \sum_{k=0}^{E}(N_k - N_k^*(v))x^k && (4.2)\\ &= \sum_{k=0}^{E}\left[R_k - \frac{R_k^*(v)}{(1-x)}\right]x^k(1-x)^{E-k}. && (4.3)\end{aligned}$$

Note that, in contrast with other measures of centrality, criticality may depend on both the *reliability rule* and the node failure probability x. For example, if node v_1 is an element of one structural motif of size k and node v_2 is an element of two structural motifs of size $k+1$, then it can be shown that to leading order, $\Delta_{v_1}(x) - \Delta_{v_2}(x) = x^k(1-2x)$ and the nodes' relative ranking changes at $x \approx 1/2$.[2] Though centrality is often considered an intrinsic property of a graph, separate from dynamics on the graph, we believe this dependence is a desirable property. In general, the criticality of an edge is a very complicated function of the overlaps among the structural motifs that contain it. However, we illustrate the concept with a simple example in which the criticality can be expressed easily. Suppose we add a structural motif of size k_0 to a graph G, and that it is edge-disjoint from any existing structural motif in the graph. Then, using Equations 3.3 and 3.13, it can be shown that the reliability of the new graph, $R'(x)$ is:

$$R'(x) = R(x)(1-x^{k_0}) + x^{k_0}. \tag{4.4}$$

Thus the criticality of any edge in the new structural motif is $R'(x) - R(x) = x^{k_0}(1 - R(x))$. Since criticality is not a constant, but a *function* of x, we may wish to choose any of several aspects of that function as our centrality measure.

[2] In [42] similar observation was reported for the edge rank method.

We consider the mean value $\langle \Delta_v \rangle \equiv \int_0^1 \Delta_v(x)dx$. It can be shown that the mean value of $R(x)$ is given by the mean value of P_k:

$$\int_0^1 R(x)dx = (E+1)^{-1} \sum_{k=0}^{E} P_k \equiv \langle P \rangle. \tag{4.5}$$

Hence the mean criticality $\langle \Delta_v \rangle$ can be expressed as a difference between the mean values of P_k and $P_k^*(v)$ taking into account, that $G - v$ has one less node than G:

$$\langle \Delta_v \rangle = \langle P \rangle - \langle P^*(v) \rangle. \tag{4.6}$$

This procedure can be generalized to assess the most important edges. We may also choose features of the coefficients P_k or N_k directly, e.g. $k_{min} - k^*_{min}$, $P_{k_{min}} - P^*_{k_{min}}(v)$. The later choice is particularly interesting because of the physical interpretation for these coefficients. It is the smallest structural motifs that determine the leading order behavior in $R(x)$, because they are associated with the term $x^{k_{min}}$. Thus $P_{k_{min}} - P^*_{k_{min}}(v)$ gives the number of the smallest structural motifs that include node v. Although, the leading order behavior in x is not necessarily the largest effect, it is undoubtedly a good place to start looking for the dynamical consequences of removing a vertex.

The difference between the coefficients for k_{min} themselves does not reflect *any* higher-order structure. We can include some of this by using the difference between reliabilities evaluated at x_{min}. In what follows we shall choose the relative difference for our measure of centrality:

$$\mu(G, v) \equiv 1 - R^*_{x_{min}}(v)/R_{x_{min}} \tag{4.7}$$

We later demonstrate an algorithm that takes into account μ as its measure of criti-

cality.

4.4.2 Estimation

Evaluating the reliability separately for every edge/node that could be removed from a large graph is an expensive computation. Fortunately, there is an easy way to estimate it in the course of estimating the reliability for G itself that involves only a small amount of additional bookkeeping.

For each value of k of interest, the algorithm suggested by Section 1.4 constructs a random sample of subgraphs with k edges and counts the number that are reliable. The bookkeeping we add is to maintain separate counts for each node v of the number of selected subgraphs that contain v, $n(v)$, and the number of those that are reliable, $r(v)$. With this information we can construct a 2x2 contingency table as in Table 4.1 for each node v. We have considered using the value of a test statistic for statistical independence, e.g. χ^2 or mutual information computed for each of these tables. Unfortunately, tests for statistical independence on these tables almost always show significant difference from the null. In part, this happens because the sample size can be extremely large; more importantly, the two factors are simply not independent – even if v is *not* critical, there is an opportunity cost incurred if it is included in a subgraph. However, the measure $\mu(G, v)$ defined above is both easily computed from these tables and easily interpreted. Thus, for each node v, we estimate $\mu(G, v) \approx \frac{r}{n} - \frac{r(v)}{n(v)}$. (Note that, given an estimate for R_k, since V and k are known, there is only one free parameter in each table. $\mu(G, V)$ is our choice of a useful representation of this parameter.)

We save ourselves the need to compute $R^*(G, x, v)$ separately for each node by accepting smaller sample sizes for our estimate of $P^*_{k_{min}}$. Indeed, since each subgraph's nodes are drawn uniformly at random from the whole graph G, we expect $n(v) \approx nk/V$. If

Table 4.1: A contingency table showing the number of reliable or unreliable graphs that do or do not include node v in a sample of size n. The values indicated in the last row and column are the marginal population totals, i.e. the values that would observed if the entire population were included in the sample. Copyright ©2013 IEEE.

	v present	v absent	
reliable	$r(v)$	$r - r(v)$	R_k
unreliable	$n(v) - r(v)$	$n - r - n(v) + r(v)$	$\binom{V}{k} - R_k$
	$\binom{V-1}{k-1}$	$\binom{V}{k-1}$	$\binom{V}{k}$

this is a problem, we can increase the sample size n to achieve sufficient precision.

There *is*, however, more useful information in the rank-ordered list of important nodes. Since we know the number of nodes in each randomly selected subgraph, k, and the total number of reliable subgraphs of this size, r, we have:

$$\sum_v r_k(v) = rk. \tag{4.8}$$

If each node appears in exactly one reliable subgraph, then there must be rk nodes with the same criticality. If there are fewer than rk nodes, the reliable subgraphs must not be node-disjoint. The number of nodes in the smallest reliable subgraphs, $n_{k_{min}}$, places constraints on the ways these smallest motifs overlap, and thus, by Equation 3.13, on their contribution to higher order terms. How best to take advantage of these constraints is an open question.

4.5 Example

We illustrate these concepts using the graph in Figure 4.3. This graph is connected and has the same number of edges as vertices. Every graph with $E = V$ is planar, i.e. can be embedded in a plane with no edge crossings, and is a forest of arbitrary-

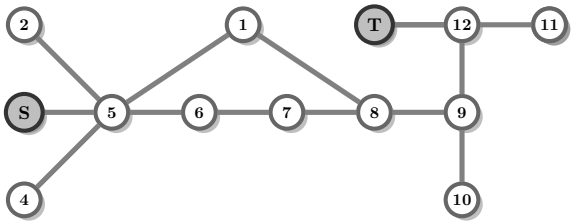


Figure 4.3: A graph with $E = V$ used to study two terminal reliability rule where nodes S and T are the source and target terminals. Copyrigh ©2013 IEEE.

sized trees rooted at the vertices of a single loop. The ST-terminal reliability for this network, for the choice of source and target depicted in Figure 4.3 is equal to: $x^7 + x^8 - x^9$.

For ST-reliability, removing any node that is not on either path from S to T has no effect on the reliability. Removing nodes that are on both paths – i.e. 5, 8, 9 or 12 – disconnects S from T, and thus leads to the largest change in reliability. Because the two branches of the loop – i.e. the "top" (5,1,8) and the "bottom" (5,6,7,8) – have different lengths, removing a node from the top has a different effect than removing one from the bottom. The reliability if a node is removed from the top branch is $R^*_{v,ST}(x) = x^8$; if an edge is removed from the bottom branch, it is $R^*_{v,ST}(x) = x^7$. Hence $\Delta_{v,ST}(x) = x^7(1 - x^2)$ or $x^8(1 - x)$, respectively; $\langle \Delta_{v,ST} \rangle \approx 0.025$ or 0.011, respectively; and the relative decrease in reliability at $x_{min} = \frac{7}{13}$ is $\mu_{ST}(v) \approx 0.569$ or 0.199. respectively.

However, for large networks, the difference in reliability induced by removing a node is mostly negligible. In this case, to apply the concept of criticality to answering question $Q2$, we define a notion of capacity for each node. Specifically, we define the generalized *capacity* of a node to be the number of structural motifs that include it. For ST-reliability, a node's capacity under this definition is the number of times it

appears in a non-self-intersecting path from S to T; for AR-0.2, it is the number of trees with exactly $0.2V$ vertices in which it appears.

Remembering that a structural motif is a *minimal* accepted subgraph, we note that any structural motif can be broken by removing any one of its nodes. Moreover, if structural motifs are not disjoint, removing a node in their overlap breaks more than one motif. Our definition of capacity properly accounts for the case of non-disjoint motifs: if a single node appears in c motifs, then removing that node breaks all c. Therefore, we define a different method to define node criticality that relies on the capacity of each node. We utilize this measure of criticality for the Colorado Springs network.

4.6 Node centrality measures

Using algorithm 4.4.2, we ranked nodes in their order of criticality. However, to be able to validate any node ranking technique, a measure that can quantifiably capture the performance of the network must be developed. For this purpose we employ network reliability. We pose that central nodes play more important roles in the network's reliability and thus their removal leads to a less reliable network. This allows us to compare different node ranking strategies. To do so, we first computed and then removed high rank nodes from the Colorado Spring network based on different centrality measures, then evaluated the reliability of the remaining network. The greater the reduction in reliability, the more critical the set of nodes and the more successful the centrality measure is. Network measures that are considered in this analysis are outlined bellow.

4.6.1 Criticality measures

Several studies have been devoted to measuring node centrality in complex networks, producing various measures of detecting and identifying the most important nodes in a network. The purpose of this section is to review a few well-known centrality measures which are later used to quantify node prominence in the spread of infectious disease.

1. The simplest measure of centrality is the **degree centrality** which is based on the number of edges between a node and all other nodes in the network. A node with high degree centrality is involved in a large number of interactions compared to other nodes in the network. Degree centrality only applies to the local structure around a node regardless of where those connections lead, i.e., it is intuitively clear that a high degree node located in a small, disconnected component of a network may not have much effect on disease propagation. [28]

2. Closeness centrality [28, 55, 76] is a measure of how close a node is to all the other nodes in the network. It is computed as the inverse sum of the shortest path distance between a node and all other nodes, therefore, the most central node based on this measure is closest to all other nodes in the network.

$$closeness_i = \frac{1}{\sum_{i \neq j} d_{i,j}} \tag{4.9}$$

where $d_{i,j}$ denotes the length of shortest path between nodes i and j.[3] From the epidemiological standpoint, this measure can be regarded as the amount of time it will take to spread the infectious disease from a node to all the other nodes.

[3]If there is no path between two nodes indicating that the graph is disconnected, the closeness is measured for each connected component separately.

3. Betweenness is another well-known measure of vertex centrality in a network, that measures the extent to which a node is involved in the interaction of the other nodes. Betweenness centrality of a node is defined as the fraction of the shortest paths between pairs of nodes in the network which pass through that node.

$$betweenness(k) = \frac{\sum_{i,j:i\neq j\neq k} \sigma_{ikj}}{\sum_{i,j:i\neq j} \sigma_{ij}} \tag{4.10}$$

where σ_{ij} denotes the number of paths between nodes i and j and σ_{ikj} is the number of such paths that contain node k. Therefore, considering the spread of infection as a flow, betweenness measures the amount of flow that travels through a given node and, in a sense is affected or controlled by that node.

4. Information centrality in another important centrality measure introduced to gauge the transmission of information between nodes in a graph. [71] It is trivial that information would traverse faster between two nodes if the path length between them is shorter, thus information is related to the reciprocal of the distance between two nodes.

$$information(i) = \frac{n}{\sum_{j\neq i} \frac{1}{I_{ij}}} \tag{4.11}$$

Where I_{ij} is the information content attributed to a path between i and j and it is inversely related to the distance of the path. If there exists more than one path between two nodes it is equal to the sum of the information of the individual paths.

5. Eigenvector centrality is another well studied measure which labels a node important if it is connected to other important nodes. [12] The eigenvector centrality of the ith node is the ith element of the largest eigenvector of the adjacency

matrix of the network.

$$A\mathbf{x} = \lambda\mathbf{x} \tag{4.12}$$

Where vector $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is the largest eigenvector of network adjacency matrix A. This measure captures the indirect effect of a node on other nodes and therefore is capable of accounting for the global influence of a node on other nodes in the network. It has been used as an influence measure in literature. [40, 29, 38]

We ranked nodes for the Colorado Springs network in Figure 4.3.2 and compared the reliability polynomials as we removed either 25 or 50 top-ranked nodes obtained based on the different centrality methods described above. Results are shown in the right panel of Figure 4.4 and 4.5, respectively. Comparing these two figures it is clear that removing more nodes reduces the reliability further. However removal of a fixed number of nodes based on different measures shows different bearing on reliability. It can be seen, that for both removal of 25 and 50 nodes, the reliability of the graph decreases more using our proposed centrality measure. Another interesting observation is the intersection of betweenness and closeness centrality in 4.5 which means that betweenness is a better measure of centrality up to near $x \approx 0.9$ while for x values larger than that closeness provides better centrality measure. Noting that x in the epidemiological context represents transmission probability of the disease, the intersection of these curves emphasize that centrality measures perform differently for different values disease transmissiblity.

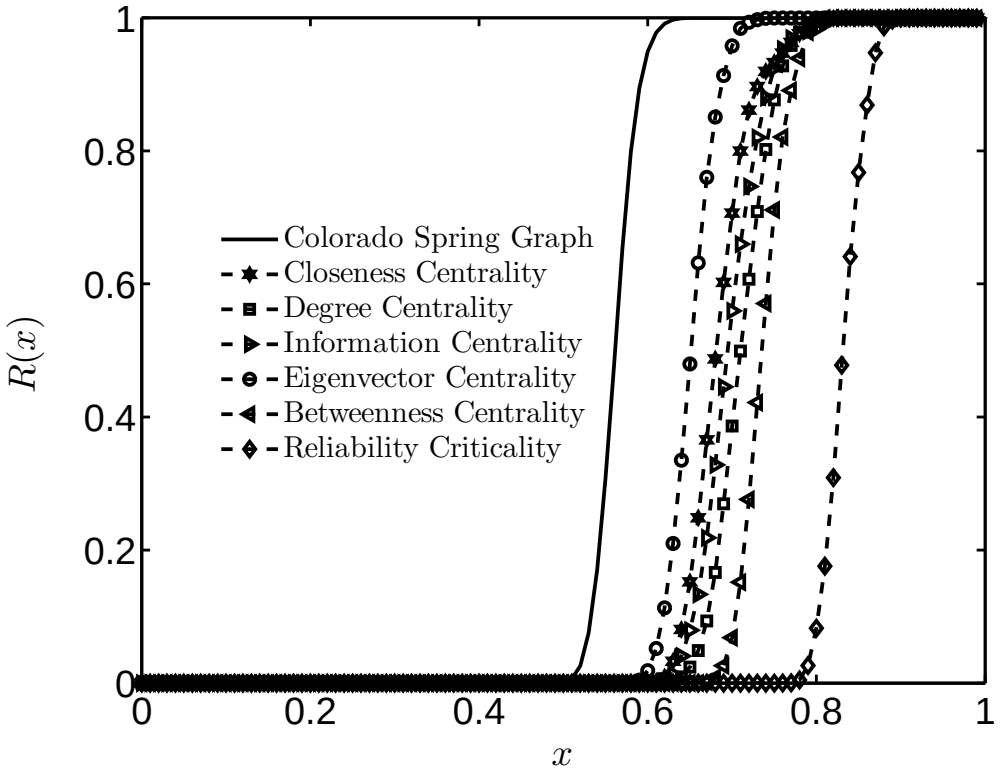


Figure 4.4: Reliability polynomial after node removal. Black curve shows the reliability of the original Colorado Springs network under AR-0.2. Black dotted lines depict the reliability of the networks after 25 high rank nodes are removed based on Closeness, Degree, Information, Eigenvector, Betweenness and Reliability centrality measures. Reliability of the network has been reduced the most based on reliability criticality measure and the least using the eigenvector centrality measure.

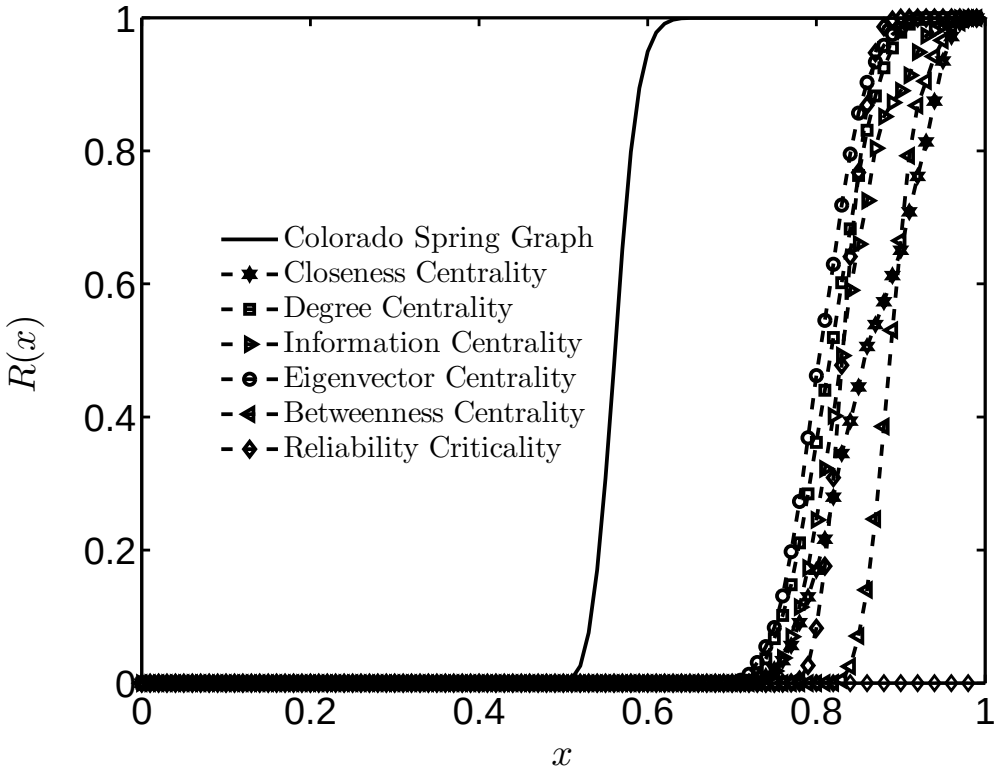


Figure 4.5: Reliability polynomial after node removal. Black curve shows the reliability of the Colorado Spring network under AR-0.2. Black dotted lines depict the reliability of the networks after 50 high rank nodes are removed based on Closeness, Degree, Information, Eigenvector, Betweenness and Reliability centrality measures. Reliability of the network after node removal based on reliability centrality measure completely decreases to zero.

4.7 Discussion

In prior work [79], we have shown that the reliability polynomial is a useful tool to study diffusive processes on networks and can be estimated efficiently. We have also illustrated how its coefficients are related to what we call structural motifs. Here we explained how to combine this estimation procedure with an understanding of the role played by structural motifs to highlight the contribution individual nodes make to reliability of networks. We showed that it is a generalization of the well-known betweenness measure of centrality, but sensitive to particular aspects of the specific dynamical process as well as the edge failure probability. Defined as a difference in reliability, it inherits the simple estimation procedures for reliability.

The ability to rank nodes in order of their contribution to dynamical phenomena supports quantitative analysis of targeted interventions in an infectious disease outbreak. We can compare the reduction in reliability achieved by the targeted intervention with both the reduction produced by various interventions and the optimal reduction, as shown in Figures 4.4 and 4.5. Moreover, the reduction can be evaluated either for a specific value of transmissibility, or in expectation across a range of possible transmissibilites.

There are many opportunities for further research in this direction, including extensions to other damage models—especially models that mix vertex and edge failures—which provide a more realistic model of disease transmission on a network. Also, defining the criticality measure to account for labeled graph, where different transmission probabilities are attributed to different classes of edges.

The results obtained here are not yet applicable to the problem of controlling an outbreak of disease in a real population. There is no practical way to determine the criticality of a contact between two people, for example. What remains is identifying

observable properties of the edges that correlate with the ranking determined here. For example, one might correlate rank with demographic labels (such as age, gender, household size) associated with vertex endpoints or the activity labels (such as work, school home) associated with the edges themselves.

Chapter V

Putting the network into network epidemiology

Contact patterns play an important role in the spread of infectious diseases in populations and various models have been developed to capture this intricate interaction. The existing works, however, have been limited to the use of structural properties to characterize the network. We argue that a proper method that can successfully estimate the behavior of an outbreak on a network, incorporates epidemiological data as a model parameter in conjunction with the properties of the network structure. We utilize the network reliability polynomial as a useful test statistic that provides approximations of the outcome of an outbreak on a network. Specifically, we show how efficient estimation of the polynomial permits characterizing and distinguishing different networks in manners that are dynamically relevant.

5.1 Introduction

Mathematical modeling for infectious disease epidemiology has focused on estimating distributions governing the natural history of a disease, such as its incubation period, severity of symptoms, and the reproductive number, R_0, defined as the average number of infections in an entirely susceptible population directly caused by a single, randomly chosen infected individual. However, there is an important difference

between R_0 and other parameters — it is determined not only by host-pathogen interactions, but also by host-host interactions [54]. Hence R_0 is not a property of a pathogen by itself, but of a pathogen in the context of a network of interactions. It is thus sensitive to the structure of this network, i.e. who can potentially transmit infection to whom. (We will use the word "contact" to indicate the potential for transmission, even if *physical* contact is not the mode of transmission.) For instance, we should replace the statement S, "The reproductive number for disease x is r", with the more meaningful statement S', "The reproductive rate for disease x on contact network N is r." Outbreak control strategies can be viewed as manipulating network structure by removing some contacts and creating others. Designing outbreak control strategies that efficiently allocate scarce resources thus requires understanding how outbreak dynamics depend on network structure.

In practice, however, we generally have only a single observation of the outbreak, which is not even a good indication of its mean, much less its variation. Even in the case of repeated outbreaks of the same disease, we cannot necessarily combine observations of outbreak information to infer its distribution, because the system is not stationary. That is, each observation is potentially taken from a different distribution. Some of the things that may change from one outbreak to another include: the strain of the pathogen, the susceptibility of the population, the population itself [6], and the outbreak control measures applied. Most important for identifying the network, though, is the problem that the contact network itself changes from one observation to another. Contact networks are not static, but change in time, location, and population surveyed.[1] The structure of a network arises from details of

[1]Contact networks co-evolve with many other things, including the prevalence of infectious disease. For example, an individual may decide not to attend school (thus breaking contacts with classmates) or to be vaccinated (thus removing a vertex and all its associated edges from the network) if the prevalence of measles is high. Our approach assumes that outbreak dynamics are faster than changes in the contact network, so it cannot address feedback loops and co-evolution. While it is obviously important to understand this entire process of co-evolution, the first step is to understand how the structure of a given network affects the dynamics of disease outbreaks.

interactions at the level of individuals, whereas we are typically interested in outbreak dynamics at the population level. One way to account for changing contact networks while simultaneously remaining focused on the population-level "forest" instead of the individual-level "trees" is to treat contact networks as random variables. That is, consider any observed contact network as a sample drawn from some probability distribution over networks. We refer to the support of this distribution as the set of realistic contact networks. [2]

Therefore, we build parameterized models for distribution over networks. The models must be *generative* — i.e. they must allow us to generate instances of networks — and they must be *probabilistic*, because we expect there to be many possible realistic networks even if all the conditions giving rise to them are held fixed. How can we determine whether the resulting model is a good one? Intuitively, we expect the model to be good if the actual process generating realistic contact networks is among the possible model processes. This requires that the model is not *mis-specified* — it does not assume relationships that are not actually present — and it is *complete* — the model parameters are in fact a sufficient set of statistics, i.e. the information contained in the modeled distribution is no less than that in the actual process.

Consider the Venn diagram in Figure 5.1, which illustrates the relationships among four sets of networks:

1. (large oval) The finite, but enormous set of all possible networks connecting N people;

2. (dashed circle) The subset of networks that could plausibly represent naturally occurring contact networks;

3. (lightly shaded oval) The set of networks that could be generated by a network

[2] We often think of the probability distribution as a uniform distribution over a restricted set of networks and speak of an instance as being in the restricted set or not, but this is just a convenient way to describe the more general case.

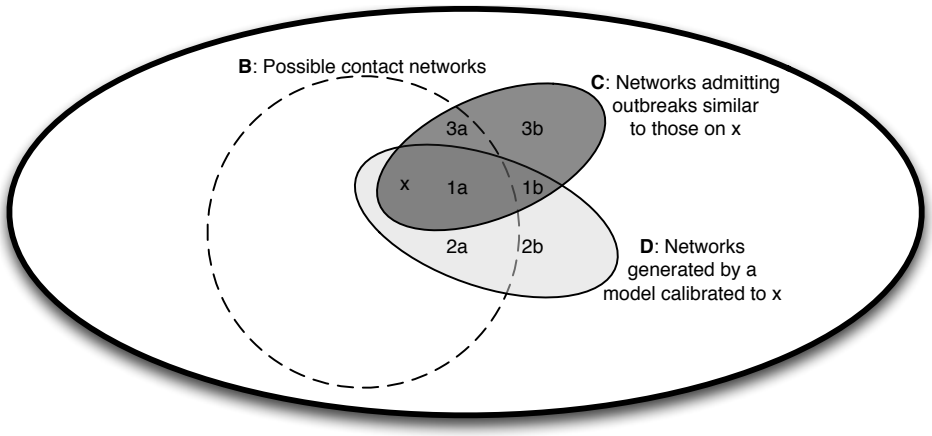


Figure 5.1: Venn diagram illustrating the possible relationships of possible social contact networks, an observed contact network (marked by the "x"), networks generated by a network model, and networks on which disease outbreaks behave in similar ways. The significance of the six labeled regions is described in the text.

model whose parameters have been calibrated to a single observed instance of a contact network; and

4. (darkly shaded oval) The set of networks on which infectious disease outbreaks are similar according to some test statistic. We shall refer to these as networks that are similar to the observed network.

Of course, the regions are not drawn to scale. Moreover, in practice sets B, C, and D do not have sharp boundaries, but are better described by a probability measure on networks. Nonetheless, we could think of the sharp boundaries in Figure 5.1 as representing a threshold probability for including a network in the set.

Consider the meaning of regions 1-3 in the figure, with respect to a single observed network x:

1. $C \cap D$: Networks *similar* to the observation x that *are* generated by our network model.

2. $\bar{C} \cap D$: Networks *dissimilar* from the observation x that *are* generated by our

network model.

3. $C \cap \bar{D}$: Networks *similar* to the observation x that *are not* generated by our network model.

Each of these regions can be further partitioned in two parts: those that (a) *are* or (b) *are not* plausible social networks. Equivalently "a" denotes the intersection of each of the above regions with B; "b", with $\bar{B}$.

What do we mean by a good or correct network model? Ideally, we might want $D = B \cap C$ or at least $B \cap D = B \cap C$.

This paper suggests a useful test statistic for studying the set C, in other words to generate networks that produce the same epidemic behavior as our target network. We argue that this suggests the need for an alternative to using constrained random graph models. One such alternative is a causal model that attempts to capture the social mechanisms driving contact network structure.

5.2 Methods for validating network models

Ideally, we would compare the distribution of realistic networks with the distribution of modeled networks, but of course, we don't know the distribution of realistic networks. Fortunately, this is a canonical problem in statistics and modeling, with a straightforward solution: estimate the probability that the observed network would have been drawn from the modeled (null) distribution.

We can make the problem a little easier by considering the questions that we are trying to answer with our network model. Generally, we need to know whether decisions about outbreak control should be made based on our understanding of outbreak dynamics on networks generated by the model. Since we don't know the

relationship between network structure and outbreak dynamics, we cannot write down the distribution induced by the model in closed form. We describe here a simple way to test the hypothesis that a candidate model of network structure is well suited for reasoning about outbreak dynamics on realistic networks.

Researchers have proposed sets of sufficient statistics for network models, often accompanied by methods for producing networks that are constrained to have particular values for these statistics but are otherwise random, which we will refer to as "constrained random networks". For example, the degree distribution (number of people with a certain number of contacts in the network) is often suggested as a sufficient statistic, and methods such as preferential attachment [8] can generate networks whose degree distributions have certain special properties. More formally, the exponential family of random graphs [63] defines a maximum entropy distribution of networks that match a variety of statistics. However, these network generation methods can be implemented efficiently only when the statistics that are constrained are relatively local ones, such as degree, assortativity, or clustering, as opposed to global structural properties such as network diameter or distribution of betweenness. Unfortunately, it is evident that outbreak features such as infection attack rate (AR) are sensitive to global structural properties, though it is not clear which ones or how the sensitivity manifests itself.

We wish to test the hypothesis that a given observed network, O, was drawn from a particular class, C, of constrained random networks that match the observed network on a proposed set of sufficient statistics. More precisely, we test whether O's structure leads to outbreaks that could, with high probability, have been drawn from those that occur on networks in C. If this probability is low, we conclude that using C to reason about the dynamics of outbreaks in real contact networks is unwise.

Given O and C, all that remains is to choose a test statistic for discriminating different

outbreak dynamics. We stress that the statistic should be directly related to important epidemiological features, i.e. there should not be a complicated chain of reasoning with many assumptions required to relate the two. The network reliability polynomial introduced by Moore and Shannon is such a statistic. [48] Although originally introduced in a rather different context, related concepts have been independently introduced to analyze contact networks for infectious disease epidemiology. [78, 56, 49] We have argued that network reliability itself provides a useful, general framework for thinking about network structure. [79, 42, 26] In this study, we describe its use for evaluating proposed equivalence classes of contact networks and demonstrate by applying to a well-known observed network.

5.3 Results for network models calibrated to friendship network

Our goal is to identify whether different network models, can generate graphs that faithfully represent the observed contact network. Our approach consists of two steps; first we generate networks calibrated to the observed network using various network models. We examine each model's goodness of fit by comparing the simulated graphs against the observed using different graph statistical metrics. A network model is accurate, if results of experiments on the simulated graphs based on that model will closely resemble those performed on the original network.

As our target social network we choose a friendship network, the faux magnolia high school, that has been studied extensively in literature. This network is derived from the National Longitudinal Study of Adolescent Health (addHealth) which is a survey study of health, demographic and relationships amongt students in 80 American high schools [35]. We studied the spread of infectious disease through this network by evaluating the distribution of the final attack rate, the day at which the number

of infected individuals reaches its maximum–peak date–and the number of infected individuals at the peak date, as depicted in Figures 5.2 and 5.3.

For network models, we compared three popular models from literature. In each of these models a set of structural properties are assumed to be essential in the formation of the network. Thus, each model claims to reproduce networks with a certain set of statistics similar to those of the observed network. Therefore, to verify each model, we compare the simulated graphs generated by that model to those of the target network, using same set of statistics used by the network model.

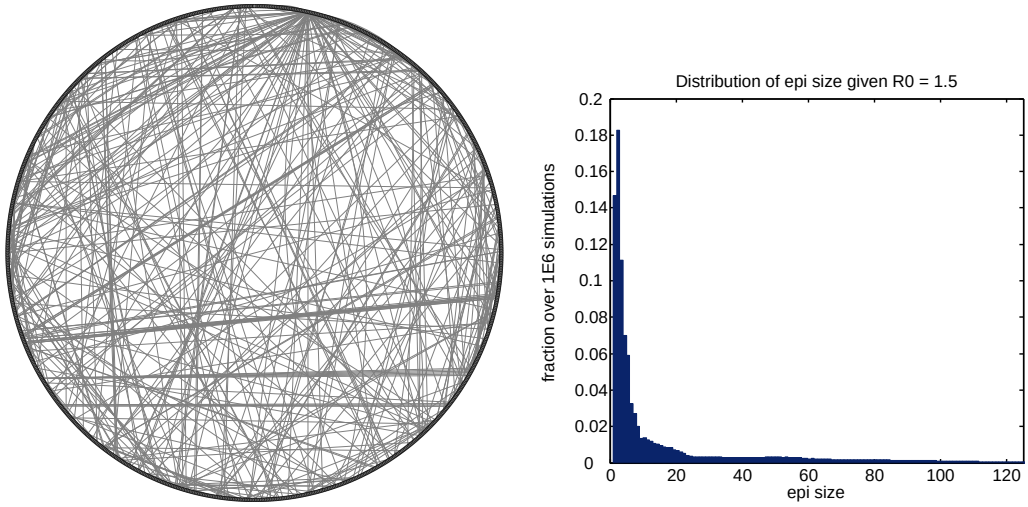


Figure 5.2: (left) Friendship network for faux magnolia high school consisting of 1461 nodes and 974 edges connecting them. It can be seen that nodes' connectivity varies by a large amount. (right) Distribution of the attack rate, for 10^6 simulation with transmissiblity $p = 0.721$ which was chosen to achieve the reproductive number $R_0 = 1.5$. Each simulation is run until no infected people remain.

One of the simplest network models is the **configuration model** that has been studied extensively since 1972. [10] A Configuration model calibrates networks to the target network by demanding the same number of contacts per node. Using this model and beginning with the friendship network, all edges are disconnected, leaving stubs (half edges) still connected to nodes. Then a graph is generated by randomly matching these stubs together. Therefore, given the degree sequence of a network,

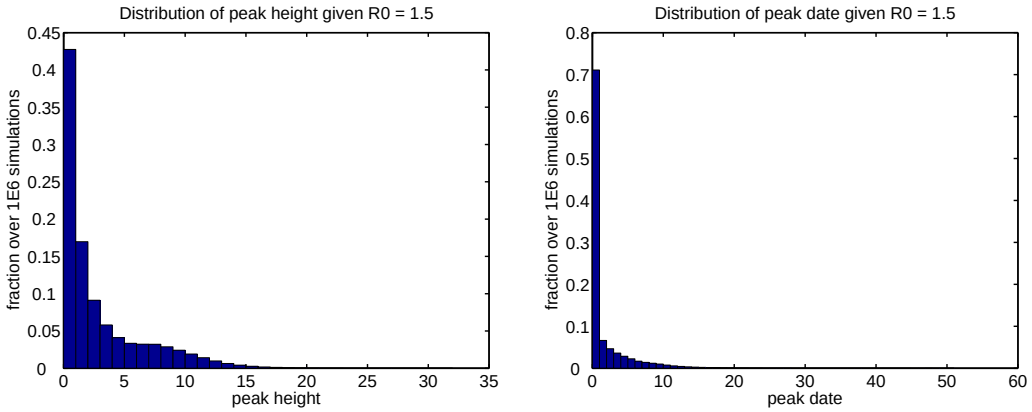


Figure 5.3: (left) The peak date of the disease season and (right) the number of infected individuals at the peak for faux magnolia high school are computed for 10^6 simulation given transmissiblity $p = 0.721$.

this model produces networks with the same degree sequence.

For the second model, we take it one step further and incorporate the clustering coefficient into the model. Clustering coefficient is a measure of the number of triangles in the network; it identifies when two nodes are connected to a third node, how likely it is for them to be connected. [9] The **random clustered model** [53, 47] is a generalization of the configuration model where, in addition to the prescribed degree distribution, the local clustering coefficient of each node is also taken into account.

The third model used in this study is the exponential random graph model (ERGM) that is undoubtedly one of the most powerful and well-studied network models.

This flexible model begins with a set of graph statistics assumed to generate the pattern of interaction observed in the network. Unlike the other two models described above, in ERGM the set of statistics are not pre-specified, a property that makes this model both complex and flexible. One can potentially choose from the large set of possible network statistics a set that is believed sufficient to generate the observed network at the same time. Identifying the sufficient set of statistics that can reproduce the observed network remains the problem to be solved. These statistics usually

include the number or density of certain structures, such as edges, triangles, stars, etc. For instance, the observation that the friend of your friend is likely to be your friend is a commonly used parameter in modeling social networks, which can be captured by counting the number of triangles. In other words, ERGM is built on the assumption that our observed network is occurring stochastically from a set of networks of a given size and certain other constraints.

Therefore, given a set of parameters, the model then assigns different probabilities to all possible networks. The goal is to find the model parameters that maximize the occurrence probability of the target network, $P(y_i)$. Furthermore, using an ERGM the probability that a random network denoted as y_i occurs is expressed as:

$$P(y_i) = \frac{e^{(-\sum_{j=1}^{k} \eta_j g_j(y_i))}}{\sum_{i \in A} e^{(-\sum_{j=1}^{k} \eta_j g_j(y_i))}} \tag{5.1}$$

Where the summation in the exponent of the exponential is over the set of the k different specified statistics and η_j is the model parameter associated with the jth statistics yet to be determined. $g_j(y_i)$ is an indicator function that determines whether the statistical property j is observed in network y_i, i.e. $g_j(y_i) \in \{0, 1\}$. The denominator is the normalizing factor ensuring that: $P \in \{0, 1\}$. The goal is then to estimate model parameters such that the observed network is attained from the model with the highest probability. [62] To define the ERGM that fits our observed network well, we should choose the set of k statistics such that the observed network can completely be distinguished from the set of all possible networks over the set of nodes.

We used the *statnet* package [34] to generate a network model which closely resembles the faux magnolia high school network. It has been suggested that a set of individual, local and network-level statistics including: density, degree and selective mixing are sufficient statistical features to simulate networks with similar structures as observed

in the friendship network. [32]

For each graph model we generated a sample of 50 networks. To validate each model we used the model statistics to compare the set of simulated networks with the target network. As can be seen in Figures 5.4 and 5.5 the statistics used in the network model—degree distribution for the configuration model and degree and clustering coefficient distribution for the random clustered model—closely match that observed in the high school network. Furthermore, for ERGM we compared along with the local structural properties of triangle participation and degree distribution the global measure of the pairwise distance between every pair of nodes. Figure 5.6 shows how well these network metrics fit the observed high school friendship network.

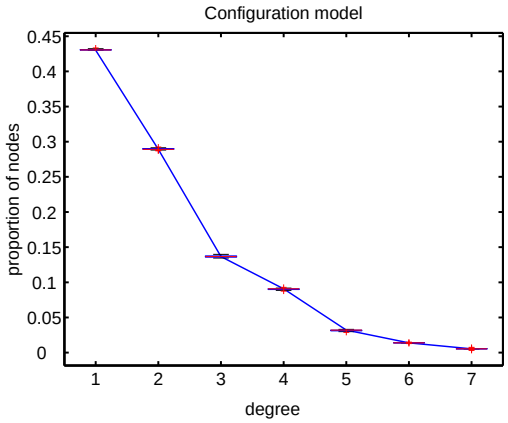


Figure 5.4: Comparing degree distribution as observed in the network set generated by configuration model with the target network. Boxes represent distribution over the set, while blue line represents the target network.

While all three models studied here can successfully reproduce similar structural statistics as they claim, it is unknown how well they can predict disease outbreaks on the network. We conjecture that in an accurate network model, epidemic outcomes on simulated networks using that model should match the observed outcome over the target network. We propose network reliability as an observable to estimate an epidemic outbreak. We further verify performance of different models with regards to

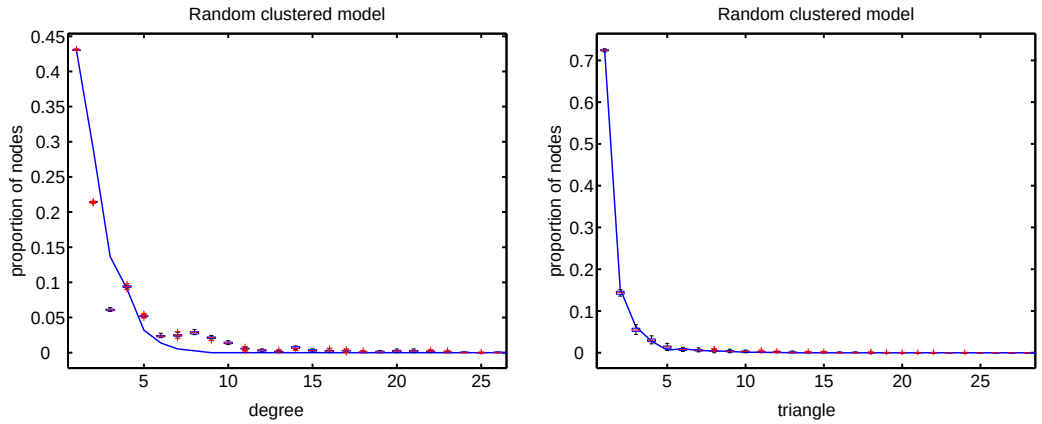


Figure 5.5: Comparing (left) degree and (right) number of triangle distributions observed in the network set generated by the random clustered model with those of the target network. The blue line represents the target network, while boxes represent distribution over the set. On each box, the red mark at center is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points within 1.5 times the interquartile range from the edge of the box; and outliers are plotted individually.

estimating the epidemic behavior through evaluation and comparison of the reliability of the graphs generated by each model with the observed friendship network.

5.3.1 Network reliability for graph characterization

Network reliability has been used as a flexible tool to study diffusive phenomena on networks, for which network-based epidemiology is a direct application.

We use the *Attack rate (AR)-α* reliability rule which indicates a subgraph is reliable if the mean component size is greater than or equal to α fraction of the initial graph. Figure 5.7 shows the reliability of the faux magnolia high school network for two values of α. The reliability of the graph increases as x, the probability of transmission, grows. In other words, when the probability of disease transmission is larger, the probability that disease affects α fraction of the population increases as well. Figure 5.7 also shows that for all values of x the probability of epidemic reaching to $\alpha = 0.1$ fraction

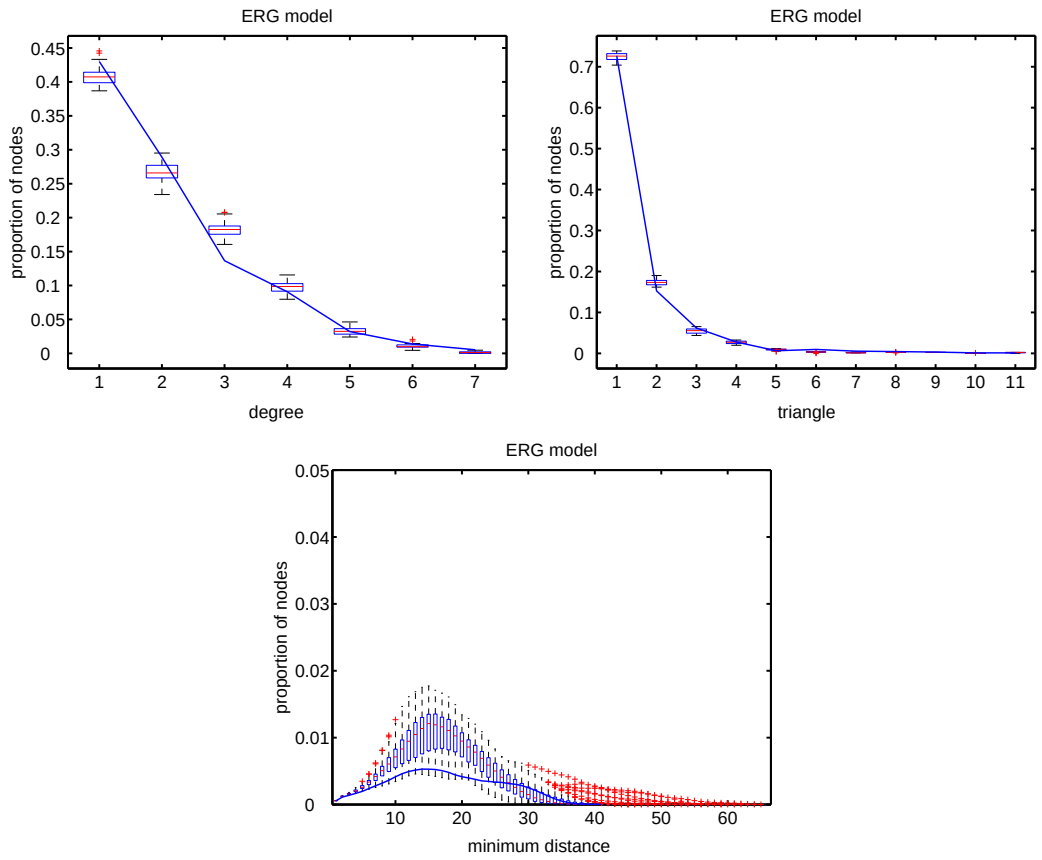


Figure 5.6: Comparing (upper left) degree distribution, (upper right) distribution of the number triangle and (lower) the distribution of the mutual minimum distance between every pair of nodes, as observed in the network set generated by ERGM with those of the target network. Boxes represent distribution over the set, while the blue line represents the target network.

of the population is higher than that of infection $\alpha = 0.2$ fraction of population. For a general introduction to network reliability, see Colbourn [18].

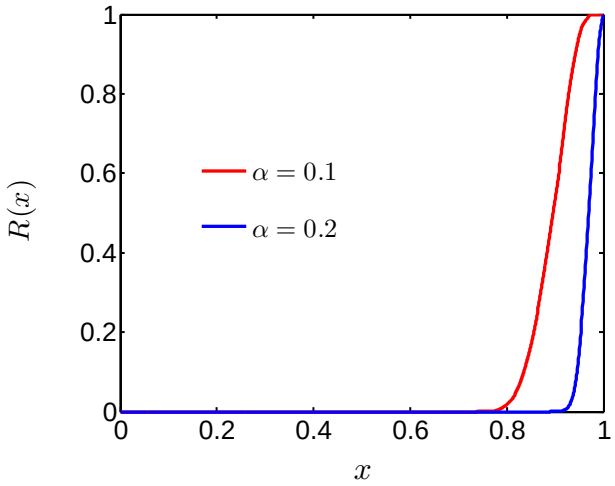


Figure 5.7: Reliability of faux magnolia high school for two values of α. The expected probability of an epidemic infecting 20% of the population is less than the probability of it infecting 0.1 fraction of the population.

We then compute the reliability for different network sets; results are shown in Figure 5.8. As can be seen, while the target network is not very reliable, networks built with different models estimate higher reliability, i.e. reliability is non-zero for smaller transmission values. The random clustered model produced the least reliable graph. Further investigation reveals that in the random clustered model, in order to obtain the desired degree and clustering coefficient distribution, more edges have been added to the network thus yielding locally denser and consequently more reliable networks. The ERGM set is the least reliable and thus most closely resembles the observed friendship network. However, despite the large variance in the reliability of the ERGM simulated set, it is still unable to fit the reliability of the friendship network.

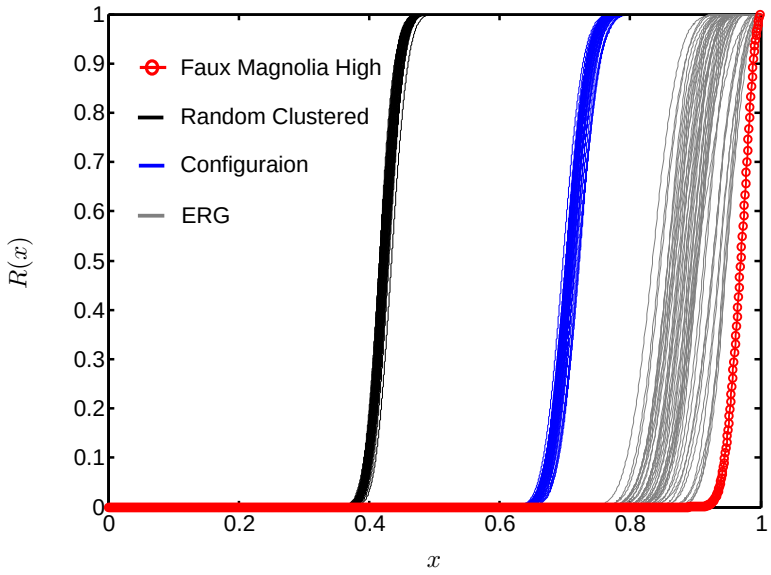


Figure 5.8: Reliability of faux magnolia high school under AR-α rule for $\alpha = 0.2$ is evaluated and compared reliability of the set of networks generated through different models. The ERGM network provide the closest estimate for the high school network.

5.4 Conclusion

Extensive research has investigated the impact of heterogeneity on disease propagation, emphasizing network-based epidemiology as a powerful alternative to replace models with homogeneous assumptions. Therefore, network models used in this approach should be able to encompass the diversity of interaction in the system. In this study, we examined several well-known network models and tested their performance in estimating an epidemic spreading on network. The results displayed in Figure 5.8 demonstrates that simulated networks based on different models, while estimating the local structural properties well, fail to capture the behavior of the disease dynamics on the network. In other words, a network cannot be characterized sufficiently without taking into account the outcome of the dynamical process taking place on it. Thus, despite the insights offered by these models, we argue that advanced models encapsulating epidemiological properties of the network as well as structural aspects

of the network are required.

We further pose that, network reliability can be utilized as a measure that captures the distribution of possible outcomes of a given disease outbreak. Thus it can be considered as a model parameter alongside the usual network structural parameters. We argue that the reliability polynomial is a natural tool for characterizing networks that considers both dynamics and structure, thus it can be used as a measure of epidemic data for designing more accurate network models.

Chapter VI

Concluding remarks

In this work we re-established the concept of network reliability to study diffusive processes on networks. The concept was first introduced by Moore and Shannon in 1950 to study the probabilistic functionality of a contact relay with faulty contacts. Numerous interpretations of this concept have emerged in different disciplines. However, due to the high computational cost, these studies has been limited to small size networks.

The reason for this complexity is two fold: the definition of the network reliability requires computation over all possible subgraphs of a graph; but, more particularly, the complex network representation employed in the measure of reliability. In contrast to many alternative measures that work with some summary statistics as network representation, including: average degree, degree distribution, etc., reliability takes the full adjacency of the network as a representation of the network structure. It thus includes the maximal information available regarding the contact pattern. This simultaneously increases the accuracy of computation done by network reliability and the complexity of this computation.

In this dissertation, starting by reviewing the common expression of reliability as it appears in the literature, we introduced a different expression that simplifies the computation. Doing so, we developed a Monte Carlo algorithm which estimates the

reliability with desired precision and eventually leads to expanding the applicability of network reliability to the set of larger, more realistic networks. This algorithm is widely appropriate–for generic network structures and generic reliability properties—and thus alleviates the need of developing algorithms for special case problems. We showed how using our algorithm we were able to study networks with millions of edges, while in the past reliability based studies had remained limited to networks with no more than 100 edges.

A set of applications of reliability is motivated by the richness of detail with which the network is represented. Reliability refocuses the question of structural effects from the local interactions between elements to global dynamical properties, suggesting new methods of analysis. This enables us to study arbitrary structural properties and their associated influence on diffusive processes on networks. We exploit this opportunity to study epidemic spreading on random networks with Erdős-Rényi and scale-free degree distribution and a range of local clustering and degree-assortative mixings. This study has direct applications in public health where designing intervention strategies emerges as a recurrent question. Reliability can verify the performance of an intervention if the intervention can be framed as a set of changes in the network's structure. For example, the simple case of isolating a single individual can be modeled by removing the node corresponding to that person from the network. In general, regardless of the intervention employed, as long as it can be interpreted as a structural change its performance can be put under scrutiny using reliability.

Furthermore, we produce a different expression for network reliability in terms of structural constituents of a network that are of functional importance which we termed structural motifs. Expressing reliability in this form enables network characterization that is tailored to the dynamical phenomena of interest. This representation expands the application of network reliability to cover a broader scope of problems, including characterizing structural differences between two networks regarding a spe-

cific dynamic. We further illustrated how different structural entities may have different bearings on the reliability of a network under different dynamical phenomena. Building on the role of structural motifs as network building blocks, we introduced a reliability based measure of criticality. This measure identifies the set of network entities that are of prominent importance in structural motifs of a network and consequently in its reliability. The direct connection of structural motifs to dynamical processes diffusing upon a network brings up a unique opportunity for gauging the role that different network entities play regarding various dynamical phenomena.

Lastly, motivated by definition of network reliability as the probabilistic measure of diffusive processes on network, we argue that it can measure the spread of epidemics on networks. Therefore, it can be employed to design network models. A network model is necessary when access to real instances of networks are impossible. However, validating what is a proper network model depends on the area of the interest. We pose network reliability as a powerful tool that can answer the questions of similarity regarding the outcome of the diffusive process upon networks.

To summarize, this dissertation was an effort to establish network reliability as a framework to study diffusive processes on large networks. More precisely the focus was on distinguishing the differences between networks regarding the specific dynamical process. Though the applications addressed were projected toward epidemiology research the unified framework can assist us to delve into problems in various disciplines.

Bibliography

[1] H.M. AboElFotoh and C.J. Colbourn. Computing 2-terminal reliability for radio-broadcast networks. *Reliability, IEEE Transactions on*, 38(5):538–555, Dec 1989.

[2] Hosam M. Aboelfotoh and Charles J. Colbourn. Series-parallel bounds for the two-terminal reliability problem. *ORSA Journal on Computing*, 1(4):209–222, 1989.

[3] Uri Alon. Network motifs: theory and experimental approachess. *Nature Reviews Genetics*, 8:450,461, 2007.

[4] David M. Auerbach, William W. Darrow, Harold W. Jaffe, and James W. Curran. Cluster of cases of the acquired immune deficiency syndrome: Patients linked by sexual contact. *The American Journal of Medicine*, 76(3):487 – 492, 1984.

[5] Michael O. Ball. Complexity of network reliability computations. *Networks*, 10(2):153–165, 1980.

[6] Shweta Bansal, Jonathan Read, Babak Pourbohloul, and Lauren Ancel Meyers. The dynamic nature of contact networks in infectious disease epidemiology. *Journal of Biological Dynamics*, 4(5):478–489, 2010. PMID: 22877143.

[7] A. L. Barabási and Z. N. Oltvai. Network Biology: Understanding the Cell's Functional Organization. *Nature Genetics*, 5:101–114, 2004.

[8] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[9] A. Barrat and M. Weigt. On the properties of small-world network models. *Eur. Phys. J. B*, 13:547, 2000.

[10] András Békéssy, P. Bekessy, and János Komlós. Asymptotic enumeration of regular matrices. *Studia Scientiarum Mathematicarum Hungarica*, 7:343–353, 1972.

[11] Ramesh Bhandari. *Survivable Networks: Algorithms for Diverse Routing*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[12] Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, pages 1170–1182, 1987.

[13] J.I. Brown, Y. Koc, and R.E. Kooij. Reliability polynomials crossing more than twice. In *Proceedings of the 3rd International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2011*, Budapest, Hungary, October 5-7, 2011, 2011.

[14] F. Buekenhout and M. Parker. The number of nets of the regular convex polytopes in dimension ≤ 4. *Disc. Math*, 186:69–94, 1998.

[15] Sergey V. Buldyrev, Roni Parshani, Gerald Paul, H. Eugene Stanley, and Shlomo Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464:1025,1028, 2009.

[16] Seth Chaiken. A combinatorial proof of the all minors matrix tree theorem. *SIAM J. ALG. DISC. METH*, 3, 1982.

[17] Shu-Chiuan Chang and Robert Shrock. Reliability polynomials and their asymptotic limits for families of graphs. *Journal of Statistical Physics*, 112(5-6):1019–1077, 2003.

[18] C. J. Colbourn. *The Combinatorics of Network Reliability*. Oxford University Press, 1987.

[19] CharlesJ. Colbourn. Combinatorial aspects of network reliability. *Annals of Operations Research*, 33(1):1–15, 1991.

[20] CharlesJ. Colbourn and DarylD. Harms. Evaluating performability: Most probable states and bounds. *Telecommunication Systems*, 2(1):275–300, 1993.

[21] D. M. Cvetkovi, M. Doob, , and H. Sachs. *Spectra of Graphs: Theory and Applications*. Wiley, 1998.

[22] Sara Y. Del Valle, J. M. Hyman, Herbert W. Hethcote, and Stephen G. Eubank. Mixing patterns between age groups in social networks. *Social Network*, 29:539,554, 2007.

[23] Ehab S. Elmallah. Algorithms for k-terminal reliability problems with node failures. *Networks*, 22:369–384, 1992.

[24] Rapoport Anatol et Yuan Y. *Some Experimental Aspects of Epidemics and Social Nets*, pages 327–348. Ablex Publishing Company, 1989.

[25] S. Eubank, H. Guclu, V. S. A. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[26] S. Eubank, M. Youssef, and Y. Khorramzadeh. Determining and understanding dynamically important differences between complex networks using reliability

induced structural motifs. In *Proceedings of the second workshop on complex networks and their applications, SITIS 2013*, Tokyo, Japan, December 2013.

[27] Stephen Eubank, Mina Youssef, and Yasamin Khorramzadeh. Using the network reliability polynomial to characterize and design networks. *Journal of Complex Networks*, 2014.

[28] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, page 215, 1978.

[29] N Friedkin. Theoretical foundations for centrality measures. *The American Journal of Sociology*, January 1991.

[30] A. Ganesh, L. Massouli, and D. Towsley. The effect of network topology on the spread of epidemic. *Proceedings of the IEEE INFOCOM Miami, FL, 05*, 2005.

[31] Jianxi Gao, Sergey V. Buldyrev, H. Eugene Stanley, and Shlomo Havlin. Networks formed from interdependent networks. *Nat. Phys.*, 8:40,48, 2012.

[32] StevenM. Goodreau, JamesA. Kitts, and Martina Morris. Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks*. *Demography*, 46(1):103–125, 2009.

[33] P. Grassberger. On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, pages 157–172, 1983.

[34] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. *statnet: Software tools for the Statistical Modeling of Network Data*. Seattle, WA, 2003.

[35] K.M. Harris, C.T. Halpern, E. Whitsel, J. Hussey, J. Tabor, P. Entzel, and J.R. Udry. The national longitudinal study of adolescent health, 2009.

[36] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Rev.*, 42(4):599–653, December 2000.

[37] M. Huxham, S. Beaney, and D. Raffaelli. Do parasites reduce the chances of triangulation in a real food web? *Oikos*, 76:284–300, 1996.

[38] Herbert Menzel James Coleman, Elihu Katz. The diffusion of an innovation among physicians. *Sociometry*, 20(4):253–270, 1957.

[39] Chin-Chia Jane, Wu-Hsien Shenb, and Yih-Wenn Laihc. Practical sequential bounds for approximating two-terminal reliability. *European Journal of Operational Research*, 195:427–441, 2009.

[40] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[41] A. Kershenbaum and R. M. Van Slyke. Recursive analysis of network reliability. *Networks*, 3(1):81–94, 1973.

[42] Yasamin Khorramzadeh, Mina Youssef, Stephen Eubank, and Shahir Mowlaei. Analyzing network reliability using structural motifs. *Phys. Rev. E*, 2015.

[43] A.S. Klovdahl, J.J. Potterat, D.E. Woodhouse, J.B. Muth, S.Q. Muth, and W.W. Darrow. Social networks and infectious disease: The colorado springs study. *Social Science and Medicine*, 38(1):79 – 88, 1994.

[44] Sander LM, Warren CP, Sokolov IM, and Koopman J Simon C. Percolation on heterogeneous networks as a model for epidemics. *Mathematical Biosciences 180*, 2002.

[45] D. Watts M. Newman, A. Barabsi. The structure and dynamics of networks. *Princeton University Press*, 2006.

[46] Achla Marathe, Bryan Lewis, Jiangzhuo Chen, and Stephen Eubank. Sensitivity of household transmission to household contact structure and size. *PLoS ONE*, 6(8):e22461, 08 2011.

[47] Joel C. Miller. Percolation and epidemics in random clustered networks. *Phys. Rev. E*, 80:020901, Aug 2009.

[48] E.F. Moore and C.E. Shannon. Reliable circuits using less reliable relays. *Journal of the Franklin Institute*, 262(3):191–208, 1956.

[49] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *European Physical Journal B*, 26:521–529, 2002.

[50] Yamir Moreno, Romualdo Pastor-satorras, and Ro Vespignani. Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B*, pages 521–529, 2002.

[51] Network Dynamics and Simulation Sciences Laboratory. Data set release 4, new river valley estimated contact network, 2013.

[52] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 286(026126), 2003.

[53] M. E. J. Newman. Random graphs with clustering. *Phys. Rev. Lett.*, 103:058701, Jul 2009.

[54] A. Nold. Heterogeneity in disease-transmission modeling. *Mathematical biosciences*, 52:227–240, 1980.

[55] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245 – 251, 2010.

[56] J. Omic P. Van Mieghem and R. Kooij. Virus spread in networks. *IEEE/ACM Transaction on Networking*, 17(1):1–14, 2009.

[57] Giuliano Andrea Pagani and Marco Aiello. The power grid as a complex network: A survey. *Physica A: Statistical Mechanics and its Applications*, 392(11):2688 – 2700, 2013.

[58] L.B. Page and J.E. Perry. Reliability polynomials and link importance in networks. *IEEE Transactions on Reliability*, pages 51–58, 1994.

[59] N. Parikh, M. Youssef, S. Swarup, and S. Eubank. Modeling the effect of transient populations on epidemics in washington dc. *Scientific Reports*, 3, 2013.

[60] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Rev. Mod. Phys.*, 87:925–979, Aug 2015.

[61] John J Potterat, Donald E Woodhouse, Richard B Rothenberg, Stephen Q Muth, William W Darrow, John B Muth, and Judith U Reynolds. Aids in colorado springs: Is there an epidemic? *AIDS*, 7(11):1517–1521, 1993.

[62] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173–191, May 2007.

[63] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173 – 191, 2007. Special Section: Advances in Exponential Random Graph (p*) Models.

[64] Jan Rong-Hong. Design of reliable networks. *Computers and Operations Research*, 20:25–34, 1993.

[65] Richard B. Rothenberg, John J. Potterat, Donald E. Woodhouse, William W. Darrow, Stephen Q. Muth, and Alden S. Klovdahl. Choosing a centrality measure: Epidemiologic correlates in the Colorado Springs study of social networks. *Social Networks*, 17(3-4):273–297, July 1995.

[66] A. Satyanarayana and Mark K. Chang. Network reliability and the factoring theorem. *Networks*, 13(1):107–120, 1983.

[67] A. Satyanarayana and R. Kevin Woodt. A linear-time algorithm for computing k-terminal reliability in series-parallel networks.

[68] Lawrence V. Snyder and Mark S. Daskin. Reliability models for facility location: The expected failure cost case. *Transportation Science*, 39:400–416, 2004.

[69] Nayla Sokhn, Richard Baltensperger, Louis-Flix Bersier, Jean Hennebert, and Ulrich Ultes-Nitschey. *Identification of Chordless Cycles in Ecological Networks*. Springer, 2013.

[70] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H. Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, Jan Timm, Sascha Mintzlaff, Claudia Abraham, Nicole Bock, Silvia Kietzmann, Astrid Goedde, Engin Toksz, Anja Droege, Sylvia Krobitsch, Bernhard Korn, Walter Birchmeier, Hans Lehrach, and Erich E. Wanker. A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122(6):957 – 968, 2005.

[71] Karen Stephenson and Marvin Zelen. Rethinking centrality: Methods and examples. *Social Networks*, 11(1):1–37, March 1989.

[72] Samarth Swarup, Stephen G. Eubank, and Madhav V. Marathe. Computational epidemiology as a challenge domain for multiagent systems. In *The Twelfth Inter-*

national Conference on Autonomous Agents and Multiagent Systems (AAMAS), Paris, France, May 5-9 2014, 2014.

[73] P. Van Mieghem, X. Ge, P. Schumm, S. Trajanovski, and H. Wang. Spectral graph analysis of modularity and assortativity. *Physical Review E*, 82(056113), 2010.

[74] Joseph A. Wald and Charles J. Colbourn. Steiner trees in probabilistic networks. *Microelectronics reliability*, 23(5), 1983.

[75] J. Wallinga. Perspective: human contact patterns and the spread of airborne infectious diseases. *Trends in Microbiology*, 7(9):372–377, September 1999.

[76] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[77] D. J. Watts and S. H. Strogatz. Collective dynamics of'small-world'networks. *Nature*, 393(6684):409–10, 1998.

[78] C. Wang Y. Wang, D. Chakrabarti and C. Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *SRDS*, pages 25–34, Florence, Italy, 2003.

[79] M. Youssef, Y. Khorramzadeh, and S. Eubank. Network reliability: The effect of local network structure on diffusive processes. *Physical Review E*, 88(052810), 2013.

Appendix A

Copyright

Figure 2.1 [used with permission]

M. Youssef, Y. Khorramzadeh, and S. Eubank. Network reliability: The effect of local network structure on diffusive processes. Physical Review E, 88(052810), 2013. Used with permission from the publisher; letter attached.

Figure 2.2 [used with permission]

M. Youssef, Y. Khorramzadeh, and S. Eubank. Network reliability: The effect of local network structure on diffusive processes. Physical Review E, 88(052810), 2013. Used with permission from the publisher; letter attached.

Figure 2.4 [used with permission]

M. Youssef, Y. Khorramzadeh, and S. Eubank. Network reliability: The effect of local network structure on diffusive processes. Physical Review E, 88(052810), 2013. Used with permission from the publisher; letter attached.

Figure 2.5 [used with permission]

M. Youssef, Y. Khorramzadeh, and S. Eubank. Network reliability: The effect of local network structure on diffusive processes. Physical Review E, 88(052810), 2013. Used with permission from the publisher; letter attached.

Figure 2.6 [used with permission]

M. Youssef, Y. Khorramzadeh, and S. Eubank. Network reliability: The effect of local network structure on diffusive processes. Physical Review E, 88(052810), 2013. Used with permission from the publisher; letter attached.

Table 2.1 [used with permission]

M. Youssef, Y. Khorramzadeh, and S. Eubank. Network reliability: The effect of local network structure on diffusive processes. Physical Review E, 88(052810), 2013. Used with permission from the publisher; letter attached.

Figure 3.1 [used with permission]

Yasamin Khorramzadeh, Mina Youssef, Stephen Eubank, and Shahir Mowlaei. Analyzing network reliability using structural motifs. Phys. Rev. E, 2015. Used with permission from the publisher; letter attached.

Figure 3.2 [used with permission]

Yasamin Khorramzadeh, Mina Youssef, Stephen Eubank, and Shahir Mowlaei. Analyzing network reliability using structural motifs. Phys. Rev. E, 2015. Used with permission from the publisher; letter attached.

Figure 3.3 [used with permission]

Yasamin Khorramzadeh, Mina Youssef, Stephen Eubank, and Shahir Mowlaei. Analyzing network reliability using structural motifs. Phys. Rev. E, 2015. Used with permission from the publisher; letter attached.

Figure 3.4 [used with permission]

Yasamin Khorramzadeh, Mina Youssef, Stephen Eubank, and Shahir Mowlaei. Analyzing network reliability using structural motifs. Phys. Rev. E, 2015. Used with permission from the publisher; letter attached.

Figure 3.5 [used with permission]

Yasamin Khorramzadeh, Mina Youssef, Stephen Eubank, and Shahir Mowlaei. Analyzing network reliability using structural motifs. Phys. Rev. E, 2015. Used with permission from the publisher; letter attached.

Table 3.1 [used with permission]

Yasamin Khorramzadeh, Mina Youssef, Stephen Eubank, and Shahir Mowlaei. Analyzing network reliability using structural motifs. Phys. Rev. E, 2015. Used with permission from the publisher; letter attached.

Table 3.2 [used with permission]

Yasamin Khorramzadeh, Mina Youssef, Stephen Eubank, and Shahir Mowlaei. Analyzing network reliability using structural motifs. Phys. Rev. E, 2015. Used with permission from the publisher; letter attached.