Analysis and Evaluation of Social Network Anomaly Detection

Meng John Zhao

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

Statistics

William H. Woodall Co-Chair

Anne R. Driscoll Co-Chair

Ronald D. Fricker

Nathaniel T. Stevens

Srijan Sengupta

August 21st, 2017

Blacksburg, VA

Keywords: change detection, Erdos-Renyi model, moving window, social network, standardization, statistical process monitoring, aggregation, DCSBM, scan method

Analysis and Evaluation of Social Network Anomaly Detection

Meng John Zhao

ABSTRACT

As social networks become more prevalent, there is significant interest in studying these network data, the focus often being on detecting anomalous events. This area of research is referred to as social network surveillance or social network change detection. While there are a variety of proposed methods suitable for different monitoring situations, two important issues have yet to be completely addressed in network surveillance literature. First, performance assessments using simulated data to evaluate the statistical performance of a particular method. Second, the study of aggregated data in social network surveillance. The research presented tackle these issues in two parts, evaluation of a popular anomaly detection method and investigation of the effects of different aggregation levels on network anomaly detection.

Analysis and Evaluation of Social Network Anomaly Detection

Meng John Zhao

GENERAL ABSTRACT

Social networks are increasingly becoming a part of our normal lives. These networks contain a wealth of information that can be immensely useful in a variety of areas, from targeting a specific audience for advertisement, to apprehending criminals, to detecting terrorist activities. The research presented focus evaluating popular methods on monitoring these social networks, and the potential information loss one might encounter when only limited information can be collected over a specific time period, we present our commendations on social network monitoring that are applicable to a wide range of scenarios as well.

Acknowledgement
I am grateful to my dissertation advisors, Dr. William H Woodall and Dr. Anne R Driscoll. For both of their guidance and support throughout the time of my dissertation research. I am thankful to Dr. Ronald D. Fricker, Dr. Srijan Sengupta, and Dr. Nathaniel Stevens for their advices and recommendations on my dissertation. Without their continued support, I could not have finished my research and my dissertation.

In addition, I would like thank Mohamed El Khouly, Adam Edwards, and Dr. Matthew Keefe for their support during the most difficult part of my graduate studies. I really appreciate their encouragement throughout this last semester. I would also like to thank Dr. Jeffrey B. Birch for believing in me when I first applied to the Statistics Department; Dr. Geoffrey Vining for the opportunity to apply the statistics knowledge I learned in real world problem solving situations.

To Christina Dillon, Betty Higginbotham, Steve Slaughter, and Debra Stover, the staff of the Virginia Tech Statistics Department. Their help have made my research possible.

Finally, to my parents, Dr. Xiaoli Glasgow and Dr. Matthew Glasgow, their support and love over these past five years have made me realize that coming to Virginia Tech and Blacksburg to pursue a doctorate degree in Statistics has been one of the best choices I made in life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

As social networks become more prevalent, there is significant interest in studying these network data, the focus often being on detecting anomalous events. This area of research is referred to as social network surveillance or social network change detection. Woodall et al. (2017), Savage et al. (2014), and Bindu et al. (2016) have reviewed a number of methods that can be used to monitor network communications for anomalous events. While there are a variety of proposed methods suitable for different monitoring situations, two important issues have yet to be completely addressed in network surveillance literature. First, performance assessments for the majority of the methods are based on case study data from existing networks. Case study data tend to only illustrate the performance of a method in one specific situation. Simulated data that are applicable to a broad array of circumstances are necessary to evaluate the statistical performance of a particular method. Second, even if the precise time of each communication instance in the network can be determined, the network data being monitored are always aggregated. Baseline communication information can vary significantly depending on the level of aggregation. This can considerably alter the performance of the anomaly detection method used. The research presented in this dissertation address these issues in two parts, an evaluation of a popular anomaly detection method and an investigation of the effects of different aggregation levels on network anomaly detection.

For the first part of this dissertation, we use simulated social network data generated based on the Erdős–Rényi model (Erdős and Rényi 1959) to assess the performance of Priebe's scan method, introduced by Priebe et al. (2005). In Zhao et al. (2017), we use synthetic data to precisely control the parameters relevant to the anomalous events. This permits us to accurately investigate the performance of Priebe's scan method. We demonstrate that the performance can be improved by some key modifications. In addition to detecting the presence of anomalous activities, locating the anomalous behavior is equally important. We recommend a modularity method developed by Girvan and Newman (2002) to identify the nodes responsible for the anomalies in the social network.

The second part of the research is to study the effect of different aggregation levels on the performance of Priebe's scan method. We take several critical characteristics, such as the aggregation level, baseline communication level in the network, and the false alarm rates into account. Our simulation study provides clear insight for practitioners on the effect of data aggregation on the performance of the Priebe et al. (2005) method. The methodology we propose can be used to study the effect of data aggregation on any surveillance method.

References

Bindu, P. V., and Thilagam, P. S. (2016). Mining Social Networks for Anomalies: Methods and Challenges. Journal of Network and Computer Applications, 68: 213-229.

Erdős, P. and Rényi, A. (1959). On Random Graphs. I. Publicationes Mathematicae, 6: 290–297.

Girvan, M. and Newman, M. E. J. (2002). Community Structure in Social and Biological Networks. Proceedings of the National Academy of Sciences, 99(12): 7821-7826.

Priebe, C. E., Conroy, J. M., Marchette, D. J., and Park, Y. (2005). Scan Statistics on Enron Graphs. Computational & Mathematical Organization Theory, 11(3): 229-247.

Savage, D., Zhang, X., Yu, X., Chou, P., and Wang, Q. (2014). Anomaly Detection in Online Social Networks. Social Networks, 39: 62-70.

Woodall, W. H., Zhao, M. J., Paynabar, K., Sparks, R., and Wilson, J. D. (2017). An overview and perspective on social network monitoring. IISE Transactions, 49(3):354-365.

Zhao, M. J., Driscoll, A. R., Fricker, R. D. Jr., Woodall, W. H., Spitzner, D. J., (2017). Performance Evaluation of Social Network Anomaly Detection Using a Moving Window Based Scan Method. 2nd revision submitted to Computational and Mathematical Organization Theory.

Chapter 2

Performance Evaluation of Social Network Anomaly Detection Using a Moving Window Based Scan Method

Meng J. Zhao, Anne R. Driscoll, Ronald D. Fricker Jr., Srijan Sengupta, Dan J. Spitzner, William H. Woodall (2017). Submitted for publication.

Abstract

Timely detection of anomalous events in networks, particularly social networks, is a problem of increasing interest and relevance. A variety of methods have been proposed for monitoring such networks, including the window based scan method proposed by Priebe et al. (2005). However, research assessing the performance of this and other methods has been sparse. In this article, we use simulated social network structures to study the performance of the Priebe et al. (2005) method. We show that it frequently cannot detect anomalous events in a network even under very simple conditions. The detection power tends to be adequate only when more than half of the social network experiences anomalous behavior or if the anomalous behavior is extreme. Both can be represented by high signal-to-noise ratios in the network. More precisely, the performance of Priebe's scan method only performs well when the signal-to-noise ratio is above 20. Simulation studies are employed to show that an improved detection rate and shortened monitoring delays can be achieved by lagging the moving window used for standardization, lowering the signaling threshold and using shorter moving windows at the initial stage of monitoring. We suggest a community detection method to be used after an anomalous event has been identified to help determine the sub-network associated with this anomalous behavior.

Keywords change detection, Erdos-Renyi model, moving window, social network, standardization, statistical process monitoring

1 Introduction

Data generated by networks are increasingly available from a variety of sources, particularly social media and other types of Internet-based applications and communications. As a result, there is significant interest in studying these types of network data, often with a focus on detecting anomalous events in social networks. Savage et al. (2014) described a number of important applications, including security monitoring, advertisement targeting, and distribution and dissemination of news. Carley (2003), McCulloh and Carley (2011), and Wei and Carley (2015) studied temporal pattern recognition and anomaly detection in social networks with some business and military applications. Savage et al. (2014), Bindu and Thilagam (2016), and Woodall et al. (2017) reviewed numerous methods that have been developed for monitoring social network data.

To make the idea of monitoring a network concrete, consider a social network consisting of communications between n individuals, where the communications data are aggregated over sequential time periods of fixed length, and the communication level between any two individuals i and j is recorded for each time period t, $t = 1, 2, \ldots$. In any given time period, the communication between individuals either may be recorded as binary data indicating that individuals i and j did or did not communicate at a specified level, where $C_t(i,j) \in \{0, 1\}$; or as count data indicating the number of communications between the two individuals, where $C_t(i,j) \in \{0, 1, 2, 3 \ldots\}$. In both cases, $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n,\ i \neq j$.

When communications are recorded as binary data, an underlying Bernoulli distribution is often assumed. When counts of the number of communications are recorded, an underlying Poisson distribution is often assumed. Clearly binary data do not provide as much information about the level of communication in a social network as count data.

For each time period t, all of the network communication information is placed in an $n \times n$ matrix C_t, referred to as an *adjacency matrix*. Using a series of historical adjacency matrices, "normal" or *baseline* levels of communication between individuals can be quantified and significant deviations from these baseline levels can then be flagged as potential anomalous events.

1.1 Types of Network Analyses

In the context of network operations, Savage et al. (2014) referred to the types of networks as static networks and dynamic networks. Broadly speaking, network surveillance only applies to dynamic networks. There are two types of analyses applied to network anomaly detection: retrospective and prospective. A fixed amount of dynamic network data is used (in a retrospective analysis) to detect anomalies in some applications. For example, Bindu et al. (2016), Ranshous et al. (2015) and Akoglu et al. (2014) reviewed methods where the main task is to identify anomalous network activities given the network history.

In contrast, streams of network data can be used to prospectively monitor the network for anomalies over time. A major challenge inherent in prospective network monitoring is that the network is generally not stationary, reflecting the changing environments both within and outside of the network, as well as the behavioral habits of the network members.

A natural idea for addressing network non-stationarity is to use a moving window of recent data as the baseline against which to look for anomalies. Moving window methods are sometimes called "scan" methods. This is the approach used in Priebe et al. (2005), one of the more widely cited papers in this literature and the focus of the present investigation. In what follows, the moving window based scan method described in Priebe et al. (2005) will be referred to as "Priebe's scan method." There are concerns associated with the use of a moving window based scan method.

For example, network anomalies that are not detected immediately result in anomalous data becoming incorporated into the moving window, making subsequent anomaly detection less likely. Another concern is monitoring delay where, depending on the design of the method, a significant data collection period might be required to establish a baseline prior to the start of monitoring. Finally, moving window based methods such as Priebe's scan method are limited to temporal based detection, and do not provide any community information on network anomalies.

1.2 Assessing Performance

While a variety of methods for monitoring social networks have been developed, there has been relatively little investigation of statistical performance. The majority of studies on social network monitoring are retrospective in which a fixed set of observed data from an existing social network or communications structure are analyzed. These types of analyses are effectively case studies and, as such, they illustrate the performance of a method in one specific situation, but they fail to assess how the method performs under a range of plausible scenarios. As Fricker (2013, p. 338) stated, "when using an actual historical set of data to evaluate performance, the data represents just one 'past reality' which, while it contains all the complications of the real world, may not reflect what might happen in the future or even what could have happened in the past under the same conditions."

Priebe et al. (2005) applied their method to a publicly available Enron email archive collected from 1998 to 2002, consisting of incidences of weekly email correspondence between 184 users, to illustrate its performance. For each pair of users in the Enron data, a "1" is recorded if at least one email was sent between the pair in a particular week, and a "0" is recorded otherwise. Thus, the network behavior

is represented by a time series of 189 adjacency matrices, each of size 184×184, where each matrix represents a single week of e-mail communications.

We agree with Woodall et al. (2017), Savage (2014) and Fricker (2013), who argued that generation of synthetic data is necessary to fully evaluate the performance of a particular method. Researchers cannot thoroughly investigate the performance of a method unless the nature and the relative timing and magnitude of anomalies are precisely known. With actual data, this information can at best only be approximately known and often it is completely unknown. In such situations, it is circular to use the method to detect what may be an anomaly only to then use the detected anomaly to assess the performance of the method. In our assessment, we identified issues such as monitoring delay due to the moving windows, inconsistency in anomaly detection when the anomalous sub-network is small, and the inability of the method to provide any information regarding the identity of the anomalous nodes. We also offer several improvements to address these issues. These include adopting a variable moving window strategy, lagging the standardization procedure, and using a more appropriate signaling threshold, as well as an algorithm that can identify the anomalous nodes in the network.

1.3 Focus of the Paper

While Priebe's scan method is frequently cited, there have been no simulation studies to evaluate its performance. Its performance properties, even in the simplest cases, are not known. In this paper we evaluate the performance of Priebe's scan method using simulated social network data based on the Erdős–Rényi model (Erdős and Rényi 1959). According to this model, when there is no anomaly present, we assume that all pairs of individuals are equally likely to communicate. To simulate anomalous events, we elevate the level of communication among some individuals within the network. We then assess Priebe's scan method according to whether it is

able to detect the increased level of communication. In addition, Montanari (2015), Hajek et al. (2015), and Hajek et al. (2017) used a signal-to-noise ratio (SNR) value as a metric to assess the network status when studying community detection. We adopt this parameter in our investigation to quantify the relative severities of the anomalous events.

A reasonable assumption is that pairwise communications in social networks at time t are affected by established patterns leading up to that time. A two-state Markov chain model, similar to the one discussed in Mousavi and Reynolds (2009), would be more appropriate for modeling this scenario. However, in order to establish baseline performance of Priebe's scan method under straightforward settings, the Erdős–Rényi model (Erdős and Rényi 1959) was chosen. Furthermore, by evaluating Priebe's scan method in this idealized, simple setting, we are able to gain insights into how the model is likely to perform under more realistic conditions and particularly whether the method has certain drawbacks, such as those resulting from contamination of the data in the moving window or a monitoring delay from having to establish moving windows. Most importantly, if the method has difficulty detecting anomalies under these idealized and simple conditions, it is reasonable to expect it will have even greater difficulties with more complex networks in more complicated scenarios.

The remainder of the paper is organized as follows. In Section 2, we introduce the necessary terminology for network analysis. In Sections 3 and 4, the methodology and our simulation strategy are reviewed. We evaluate the performance of the method for detecting anomalous events in social networks using the appropriate metrics in Section 5. Our strategies for improving the method's performance are presented in Section 6. In Section 7, we explain the approach for identifying potential anomalous sub-networks. Finally, a brief summary of our findings is presented in Section 8.

2 Social Network Terminology and Model

We define social network terminology in this section for the general reader who may be unfamiliar with both the terminology and notation. A social network $N = (V, E)$ consists of a set of *nodes* V and a set of *edges* E. In a social network, a node usually represents an individual or the origin/destination of a communication. When the nodes in a social network contain information that makes them readily identifiable, such as a name or an e-mail address, the social network is said to be *labeled*. If no identifying information is available, it is an *unlabeled* social network. The edges in the social network specify the level of communication between each pair of nodes. We denote the number of nodes and edges as $|V|$ and $|E|$, respectively. The number of nodes, $|V|$, is referred to as the *order* of the social network. The number of edges, $|E|$, is referred to as the *size* of the social network.

Assuming nodes do not change, the communication levels between all pairs of nodes at time t for a social network $N_t = (V, E_t)$ is fully summarized in an adjacency matrix C_t of dimension $|V| \times |V|$. The matrices C_t are available at the current time ($t = T$) and all previous time points, $t = 1, 2, \ldots, \mathrm{T} - 1$, each representing a snapshot of the social network structure at one of those specific times. The entry of C_t in row i and column j, denoted by $C_t(i, j)$, is a record of the communication level between node i to node j. Thus, the direction of communication is indicated by the upper right or lower left half, of off-diagonal entries where $C_t(i, j)$ lies. Values on the diagonal are zeros. Often, the direction of communication is either not measured or unknown, in which case we have $C_t(i, j) = C_t(j, i)$, so that C_t is a symmetric matrix. When only the presence or absence of a specified level of communication is measured, each $C_t(i, j)$ is restricted to the values 1 (presence) or 0 (absence). When the total number of communications is measured,

each $C_t(i,j)$ element can take any non-negative integer value. More precisely, under the Erdős–Rényi model,

$$P[C_t(i,j) = x] = p^x(1-p)^{1-x}$$

where

$$x = 0, 1.$$

Each instance of communication between any two nodes in a social network occurs with the same probability p. When there is no anomaly, we have $p = p_0$. When a relatively smaller sub-network with k nodes becomes active and experiences elevated levels of communication among its member nodes, this is considered as an anomalous event within the social network. The elevated probability of communication is denoted as p_1, where $p_1 > p_0$, with the rest of the network still communicating with the probability p_0.

The k^{th} *neighborhood* of node i contains all nodes and all edges among those nodes that are within k edges of node $i, k = 1,2,3, \ldots$. The *size* of the k^{th} neighborhood of node i is the count of all edges between nodes in the k^{th} neighborhood of node i. At time t, this quantity is denoted by $O_{t,i}^k$. We let the quantity $O_{t,i}^0$ represents the *degree* of node i , i.e., the number of direct communications to or from node i at time t.

In addition, the signal-to-noise value, or the SNR value, denoted as λ can be used to quantify the severity of anomalous events, where

$$\lambda = \frac{K^2(p_1 - p_0)^2}{(|V| - K)p_0}, \quad (1)$$

the quantity K is the is the number of nodes in the sub-network. The calculation for the numerator in Equation (1) is akin to the square of the expected increase in communication among the nodes within the anomalous sub-network. The calculation for the denominator on the other hand, is the expected communication among the non-anomalous nodes in the network. Large λ parameter values indicate strong deviations within the anomalous sub-network from normal network behavior.

Figure 1 illustrates these concepts with a simple example. The nodes enclosed in the circle in Figure 1 represent the 1st neighborhood of node 7. If this is at time t, the degree of node 7 is 2 ($O_{t,7}^{0} = 2$) because there are direct communications (edges) between node 7 and nodes 3, and 5. The size of the first neighborhood of node 7 is 3 ($O_{t,7}^{1} = 3$), because an additional communication instance (edge) exists between nodes 3 and 5. In a similar fashion, the size of the 2nd order neighborhood is $O_{t,7}^{2} = 9$.

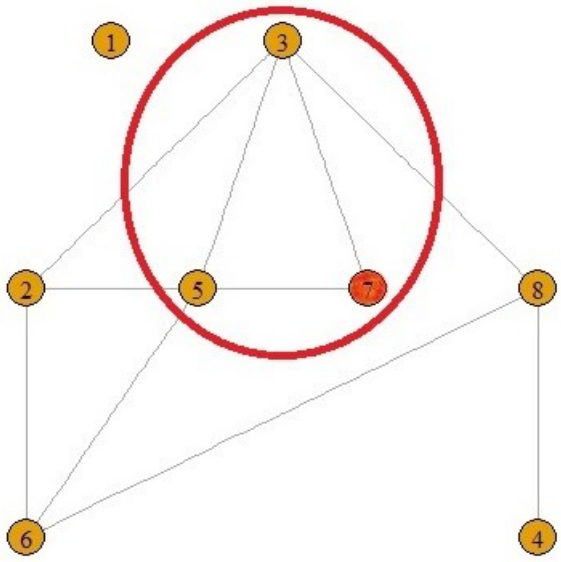

Figure 1. Network illustration. The circle indicates the first neighborhood of node 7.

3 Moving Window Based Scan Method

Priebe's scan method is based on the standardized maxima of degree, $O_{t,i}^0$, and first two neighborhood size statistics, $O_{t,i}^1$ and $O_{t,i}^2$, across all nodes. The moving window based scan statistics are calculated using a two-step standardization procedure within moving windows of width w, which Priebe et al. (2005) set to $w = 20$. At time t, $t > w$, the first standardization $O_{t,i}^{k\,*}$ is calculated as

$$O_{t,i}^{k\,*} = \frac{O_{t,i}^k - avg(O_{t,i}^k)}{\max(sd(O_{t,i}^k), 1)}, \tag{2}$$

where

$$avg(O_{t,i}^k) = \frac{1}{w}\sum_{j=1}^{w} O_{t-j,i}^k \text{ , and}$$

$$sd(O_{t,i}^k) = \left[\frac{1}{w-1}\sum_{j=1}^{w}\{O_{t-j,i}^k - avg(O_{t,i}^k)\}^2\right]^{\frac{1}{2}},$$

for $k = 0, 1, 2$ and $i = 1, 2, \ldots, |V|$. Thus, the standardized values $O_{t,i}^{k\,*}$ are calculated using information in the moving window of adjacency matrices, where $avg(O_{t,i}^k)$ and $sd(O_{t,i}^k)$ are the sample mean and standard deviation for node i calculated using the data from the previous *w* time periods. The lower bound of 1 for the denominator in Equation (2) is used to prevent detection of relatively small changes in the social network.

At time t, in the first step of standardization, the statistics $O_{t,i}^{0\,*}$, $O_{t,i}^{1\,*}$, and $O_{t,i}^{2\,*}$ are calculated from the $O_{t,i}^0$, $O_{t,i}^1$, and $O_{t,i}^2$ values, using Equation (2), for all nodes, $i = 1, \ldots, |V|$. This standardization is done at times $t = w + 1, w + 2, \ldots$. In

the second step of standardization, the statistics M_t^0, M_t^1, and M_t^2 are calculated as the respective maxima of $O_{t,i}^{0*}$, $O_{t,i}^{1*}$, and $O_{t,i}^{2*}$: $M_t^k = \max\{O_{t,1}^{k*}, O_{t,2}^{k*}, \ldots, O_{t,|V|}^{k*}\}$. These statistics reflect the highest magnitude of activity change among all nodes in the social network at time t. The statistics M_t^{0*}, M_t^{1*}, and M_t^{2*} are then calculated from the M_t^0, M_t^1, and M_t^2 values using the following equations in a second round of standardization:

$$M_t^{k*} = \frac{M_t^k - avg(M_t^k)}{\max(sd(M_t^k), 1)}, \tag{3}$$

where

$$avg(M_t^k) = \frac{1}{w}\sum_{j=1}^{w} M_{t-j}^k \ , and$$

$$sd(M_t^k) = \left[\frac{1}{w-1}\sum_{j=1}^{w}\{M_{t-j}^k - avg(M_t^k)\}^2\right]^{\frac{1}{2}},$$

for $k = 0, 1, 2$. Again, the lower bound of 1 is used in the denominator in Equation (3). The second standardization is done at times $t = 2w + 1, 2w + 2, \ldots$. It is important to note that this method requires an initial $2w$ time periods of data, which means that monitoring cannot begin until the process has been observed for a potentially extensive "pre-monitoring" period of time.

The social network is monitored over time using statistics M_t^{0*}, M_t^{1*}, and M_t^{2*}. The rule proposed in Priebe et al. (2005) is to signal the possible presence of an anomaly when any one or more of the three standardized maxima statistics calculated from Equation (3) exceeds the threshold value of 5. That is, the method signals that an anomaly may be present in the social network at time *t* if $M_t^{0*} > 5$, $M_t^{1*} > 5$ or if $M_t^{2*} > 5$.

4 Simulations

As discussed in Section 1, assessing performance using simulated social networks is key to understanding how and when a particular method performs well. In particular, simulations allow explicit control of all aspects of the network including size and complexity and, most importantly, the existence, type, and timing of anomalies. Fricker (2011) noted that proper simulation studies can facilitate evaluating algorithms across many scenarios, eliminate unneeded or distracting complexities, allow clean and clear comparisons of algorithms, and make it easier to obtain generalizable conclusions and results. In the quality control field and to some extent the field of bio-surveillance, the use of simulation is routinely used to assess and compare performance among anomaly detection methods. See, for example, Fricker, Hegler, and Dunfree (2008), and Fricker, Knitt, and Hu (2008).

Our social network simulations are based on the Erdős–Rényi model (Erdős and Rényi 1959). Under this model, each instance of communication between any two nodes in a social network occurs with the same baseline value of p_0, so that at time t when no anomalies are present $C_t(i,j)$ is a Bernoulli random variable for all i and j, $i \neq j$. Figure 2 (a) shows a heatmap representing the adjacency matrix based on a randomly generated social network of order $|V| = 100$ with $p_0 = 0.1$.

Other models for communication probabilities between individuals in a social network are possible, such as a logistic regression model used in Azarnoush et al. (2016), or the degree-corrected stochastic block model discussed in Karrar and Newman (2011). Both of these models are more flexible than the Erdős–Rényi model, and are more capable of capturing realistic social network behavior. The Erdős–Rényi model is advantageous for its simplicity and ease of interpretation.

When an anomalous event is present within the social network, the probability of communication increases from p_0 to p_1. Figures 2 (a) and 2 (b) illustrate this idea.

We simulated a network of the order $|V| = 100$. Here the probabilities of communication among nodes 25 to 55 were shifted to $p_1 = 0.9$ (and it remains at $p_0 = 0.1$ for all other nodes), in which the anomaly is clearly visible in Figure 2(b) when compared to Figure 2 (a).

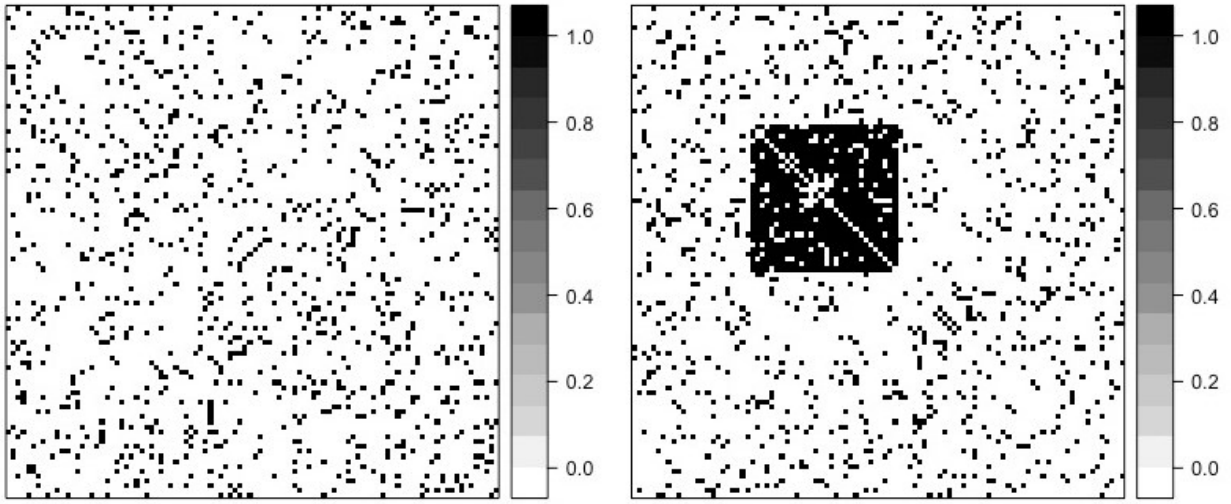

Figure 2. (a) (Left) Heat-map representing the adjacency matrix of a simulated social network of order $|V| = 100$ using baseline value $p_0 = 0.1$.
(b) (Right) Heat-map after the probability of communication for nodes 25 to 55 increased to $p_1 = 0.9$.

This generally results in increases in $O_{t,i}^0$, $O_{t,i}^1$ and $O_{t,i}^2$ of the member nodes within the sub-network. Any increases in communication levels should also be reflected by the standardized maxima statistics, M_t^{0*}, M_t^{1*}, and M_t^{2*}. Ideally, the method will promptly signal the presence of an anomaly by detecting any increases in the probability of communication within a sub-network. However, because Priebe's scan method incorporates two successive standardizations of the data, we show in the next section the value of p_1 must often be much greater than the baseline value p_0 in order to produce large enough values of M_t^{0*}, M_t^{1*}, and M_t^{2*} for detection to occur.

5 Performance Evaluation

In this section we evaluate the performance of Priebe's scan method on a variety of simulated social networks. In doing so, we vary the network order $|V|$, the baseline probability of communication, p_0, the fraction of the network with increased communication levels among member nodes (the "anomalous sub-network"), and the increased probability of communication among the members of the anomalous sub-network, p_1. Note that for a given network of order $|V|$ and a baseline probability of communication p_0, the expected degree of node i is $(|V|-1)p_0$. Given that the degree of node i is proportional to the order, $|V|$, in the simulations we adjusted the values of p_0 depending on the order of the social network.

5.1 Illustration of Priebe et al. (2005) Scan Method

To begin, here we illustrate the performance of Priebe's scan method on a social network of order $|V| = 100$. To do so, we repeatedly generated sequences of 60 adjacency matrices C_t, $t = 1, 2, \ldots, 60$, which one might think of as, say, weekly network communications data. Prior to $t = 51$, the probability of any node i in the social network communicating with any other node j is set to $p_0 = 0.02$ to simulate the baseline level of communication within the network. At time $t = 51$ and continuing until time t=60, the probability of communication increases to $p_1 = 0.18$ within a randomly selected anomalous sub-network of 10 nodes. The SNR during this period is 1.42.

The results for one simulation are shown in Figure 3. Figures 3 (a) and 3 (b) show time series plots for the degree $O_{t,i}^0$ and the standardized degree $O_{t,i}^{0*}$ for the anomalous nodes, respectively, for a 10% anomalous sub-network. The maximum of the standardized degree M_t^0 and the standardized maximum M_t^{0*} of all nodes are shown in Figures 3 (c) and 3 (d), respectively. For this simulation, Figure 3 shows

that evidence of an anomalous sub-network does not show up particularly strongly in plots (b), (c), and (d) on or after time $t = 51$, even though their probability of communication is nine times higher than the baseline probability. The standardized maximum $M_t^{0\,*}$ statistics in Figure 3 (d) does not exceed the threshold of 5 and thus a signal would not be generated. In practice, the size of the increase in the probability of communication of interest would vary by application and may be much larger than that illustrated here.

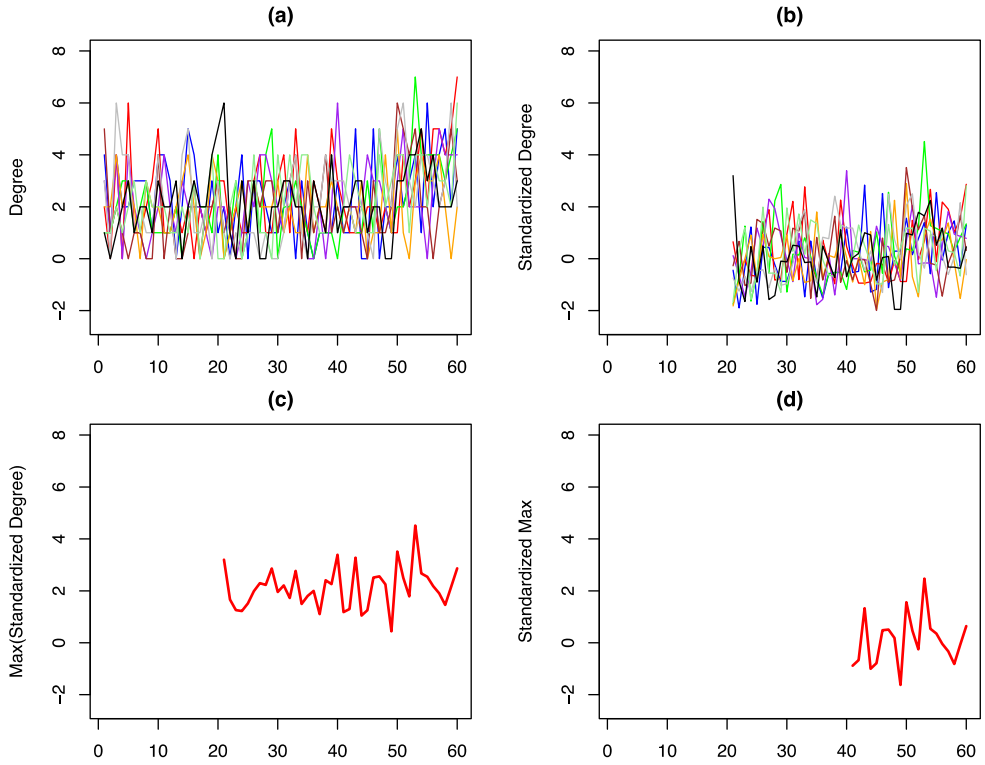

Figure 3. Time series plots of (a) $O_{t,i}^0$, (b) $O_{t,i}^{0\,*}$, (c) M_t^0, and (d) $M_t^{0\,*}$ for a 10% anomalous sub-network in a social network of order $|V| = 100$, with baseline $p_0 = 0.02$ and $p_1 = 0.18$ for $t \geq 51$. Figures 3 (a) and (b) only show results for the anomalous nodes.

Of course, when the anomalous sub-network constitutes a higher percentage of the entire social network, or p_1 is increased, then it is easier to detect the network anomaly. For example, Figure 4 shows a second simulation with the baseline

probability of communications set at $p_0 = 0.02$ and, beginning at $t = 51$, the anomalous sub-network communication probability set to $p_1 = 0.18$. What is different here is that the anomalous sub-network percentage is now 30% (30 nodes out of 100), compared to 10% for the simulation relevant to Figure 3, and thus a higher percentage of network communication is anomalous. We see in Figure 4 that, in spite of the clearly visible increases in the degree $O_{t,i}^0$ at time $t = 51$ and beyond, in the standardized degree $O_{t,i}^{0*}$ and in the maximum of the standardized degree M_t^0 at time $t = 51$, the standardized maximum M_t^{0*} statistic only briefly spikes up above the threshold of 5 at time $t = 51$ and then quickly returns to its nominal level.

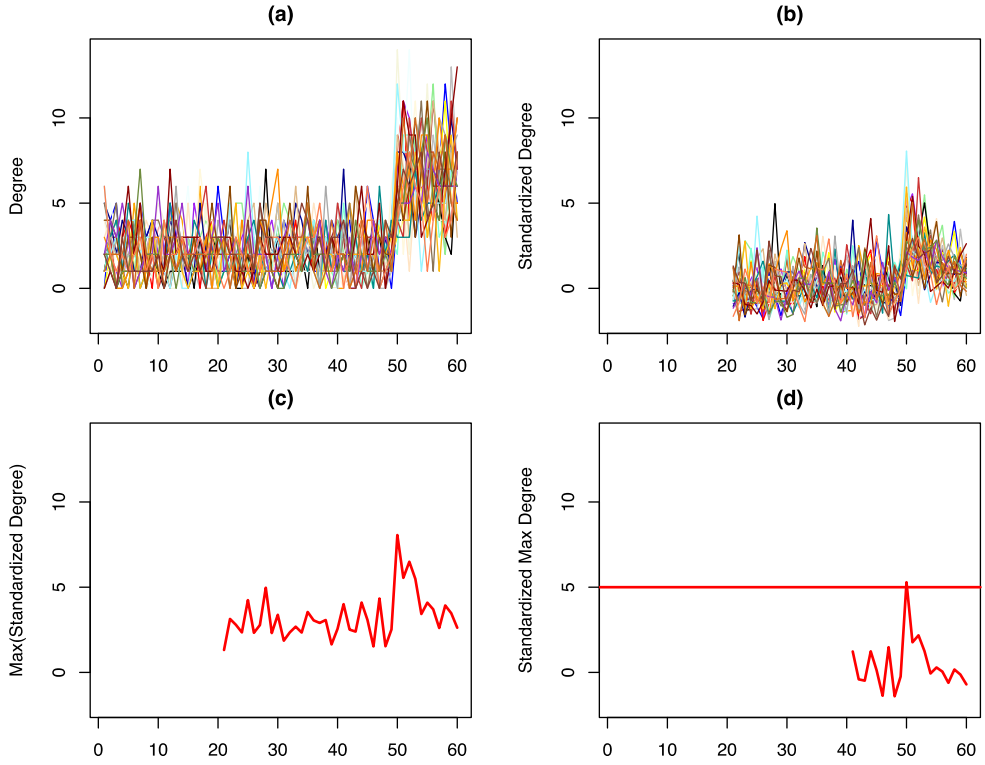

Figure 4. Time series plots of (a) $O_{t,i}^0$, $(b)O_{t,i}^*$, (c) M_t^0, and (d) M_t^{0*} for a 30% anomalous sub-network in a social network of order $|V| = 100$ with baseline $p_0 = 0.02$ and $p_1 = 0.18$ for $t \geq 51$.

Furthermore, note that in the Figure 4 (b), (c), and (d) plots that $O_{t,i}^{0*}$, M_t^0, and M_t^{0*} values start to decrease after time $t = 51$. This is visually evident, where the spike at time $t = 51$ is prominent and then each of the time series clearly and quickly begins to decrease back to its nominal level. This occurs because after time $t = 51$ the anomalous data starts being incorporated into the window upon which the most recent observation is standardized. Thus, as time progresses, the anomalous data starts to become the new baseline data and so the standardized values return to show baseline behavior. What this suggests is that if Priebe's scan method or any other similar moving window based method initially fails to detect an anomaly then it becomes less and less likely to signal.

In Figures 3 and 4 we only considered the standardized maximum of the degree, M_t^{0*},which begs the question of whether the standardized maxima of the sizes of the first and second neighborhoods perform any better for anomaly detection. Figure 5 is a time series plot for all three standardized maxima M_t^{0*}, M_t^{1*}, and M_t^{2*} for the 10% anomalous sub-network with the same simulation settings as those for Figure 3. For this particular simulation run, none of the three statistics detected any anomalous event in the social network where, at $t = 51$, there are no visible deviations from the baseline and none of the three maxima exceed the signal threshold. That is, Priebe's scan method failed to detect the anomaly.

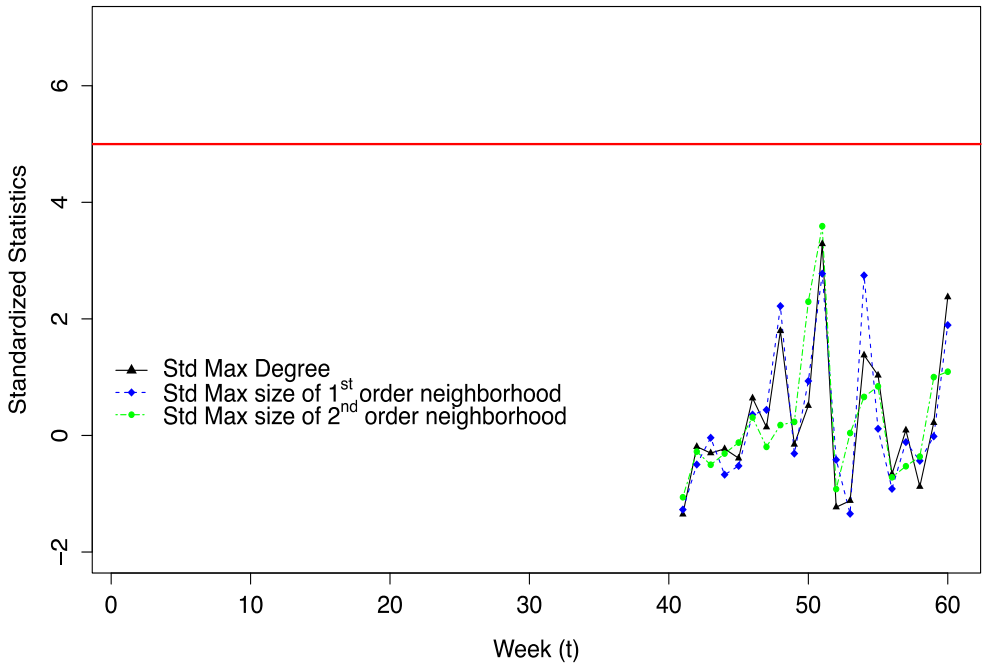

Figure 5. Time series plots of $M_t^{0\,*}$, $M_t^{1\,*}$ and $M_t^{2\,*}$ for a 10% anomalous sub-network in a social network of order $|V| = 100$, with baseline $p_0 = 0.02$ and $p_1 = 0.18$ for $t \geq 51$. The red line is the signaling threshold of 5.

Similarly, we plotted the standardized maxima for a 30% anomalous sub-network. The SNR increased substantially to 16.46, with the increase in the number of anomalous nodes in the sub-network. For the simulation result shown in Figure 6, we see that $M_t^{1\,*}$ and $M_t^{2\,*}$ values exceeded the threshold, but the $M_t^{0\,*}$ values did not. Priebe's scan method detected the anomaly.

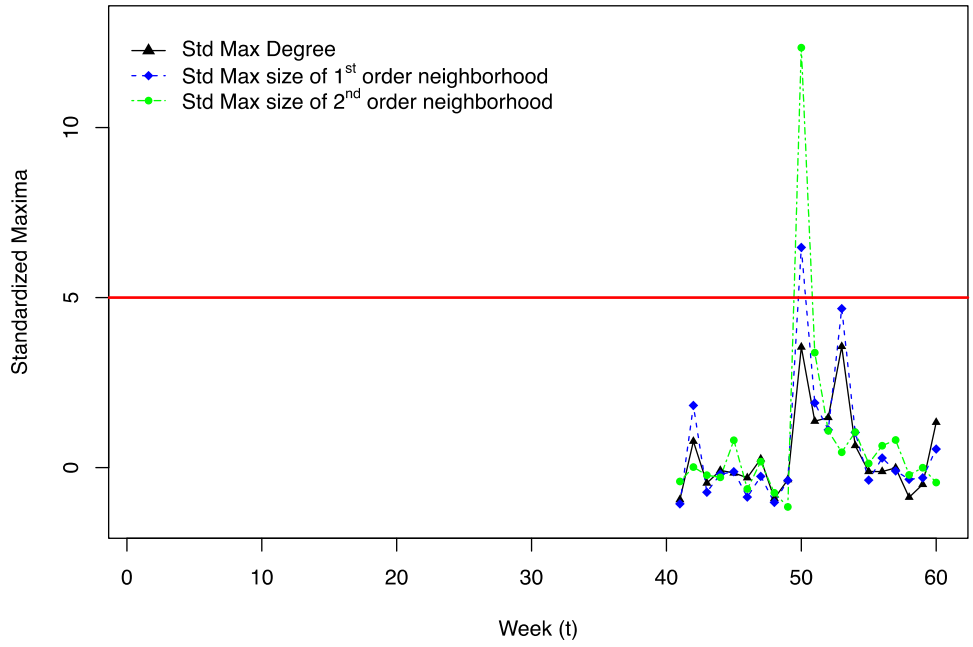

Figure 6. Time series plots of $M_t^{0\,*}$, $M_t^{1\,*}$ and $M_t^{2\,*}$ for a 30% anomalous sub-network in a social network of order $|V| = 100$ with baseline $p_0 = 0.02$ and $p_1 = 0.18$ for time $t \geq 51$. The red line is the signaling threshold of 5.

These two simulations suggest that the Priebe's scan method may not be particularly effective at detecting anomalous events in social networks when the order of the anomalous sub-network is a small percentage of $|V|$ and p_1 remains fairly small. We also observe that as the order of the anomalous sub-network increases, the SNR rises significantly, making detection more likely. For any given social network structure, one is likely to encounter cases where the vast majority of nodes in a social network continue to exhibit baseline behavior throughout the monitoring period. In other words, it is much more realistic that only a relatively small number of nodes display anomalous activity than the entire social network. Given a large social network, this relative minority could be significantly less than 10%. In the next section we explore the scan method's performance more broadly.

5.2 Performance and Detection Power

The examples in Section 5.1 suggest, perhaps unsurprisingly, that the size of the anomalous sub-network affects the ability of Priebe's scan method to detect anomalous events. However, these examples were for only a few simulation runs, so in this section we summarize the results of an extensive simulation study that evaluates the performance of Priebe's scan method under varying sub-network orders and different shifts in the probability of communication.

In particular, we conducted simulations over a wide range of anomaly probabilities, p_1, and recorded the percentage of times that at least one of the three standardized maxima statistics, $M_t^{0\,*}$, $M_t^{1\,*}$, and $M_t^{2\,*}$ exceeded the signal threshold, excluding any false signals (i.e., signals that occurred prior to time $t = 51$). We refer to this percentage as the detection power of the moving window based scan method.

In the simulations, we generated a sequence of adjacency matrices representing the baseline period from time $t = 1$ to time $t = 50$, as described in the previous section, for a network of some specified order $|V|$ and some specified value of p_0. Then from time $t = 51$ to time $t = 60$ we simulated an anomaly in the network for a specified p_1 within a chosen sub-network of a set number of nodes, where the number was set as a fraction of the total number of nodes. One thousand simulations were run for each of the scenarios listed in Table 1.

We evaluated networks configured at various configurations for $|V|$, p_0, p_1, and the relative percentage order of the anomalous sub-network in relation to the entire network, K/n. The configurations are shown in Table 1. We adjusted the p_0 value for different social network orders to keep the expected degree of any single node approximately constant and then evaluated detection performance for a variety of p_1 values. We included $p_1 = 1$ in the simulation to examine the detection power of the method in the most extreme case where all nodes within the anomalous sub-

network are certain to communicate with all other nodes within the sub-network, forming a clique. The false alarm rates of Priebe's scan method are discussed in Section 6.2.

Table 1. Simulation settings for social network scenarios.

$\|V\|$	p_0	K/n	p_1	SNR
20	0.1	10%, 20%, 30%, 50%	$0.5, 0.75, 0.9, 0.91 - 0.99, 1$	$0.36 - 81$
50	0.04	10%, 20%, 30%, 50%	$0.2, 0.3, 0.35 - 0.95, 1$	$0.36 - 576$
100	0.02	10%, 20%, 30%, 50%	$0.1, 0.15, 0.2 - 0.95, 1$	$0.36 - 2401$
200	0.01	10%, 20%, 30%, 50%	$0.05, 0.075, 0.09, 0.1, 0.15, 0.2 - 0.9, 1$	$0.36 - 9801$

Figure 7 summarizes the detection power for a social network of order $|V| = 100$ with $p_0 = 0.02$. In it we plotted detection power for 10%, 30% and 50% sub-networks versus p_1. It shows that Priebe's scan method is able to detect an anomalous event virtually 100% of the time for $0.02 < p_1 \leq 1$ when the anomalous sub-network constitutes half of the entire network. In contrast, the detection power decreases as the size of the anomalous sub-network decreases, and where we see that detecting an anomalous event is much more difficult when the anomalous sub-network is only 10% of the network, particularly for smaller values of p_1. Additionally, it is particularly interesting that the detection power is similar for both 30% and 50% anomalous sub-networks when the SNRs are the same. Using Equation (1), p_1 values of $0.63, 0.2$, and 0.11 result in a SNR value of 20.5 for 10%, 30% and 50% sub-networks respectively, where the detection power for all three sub-network orders is close to 100%.

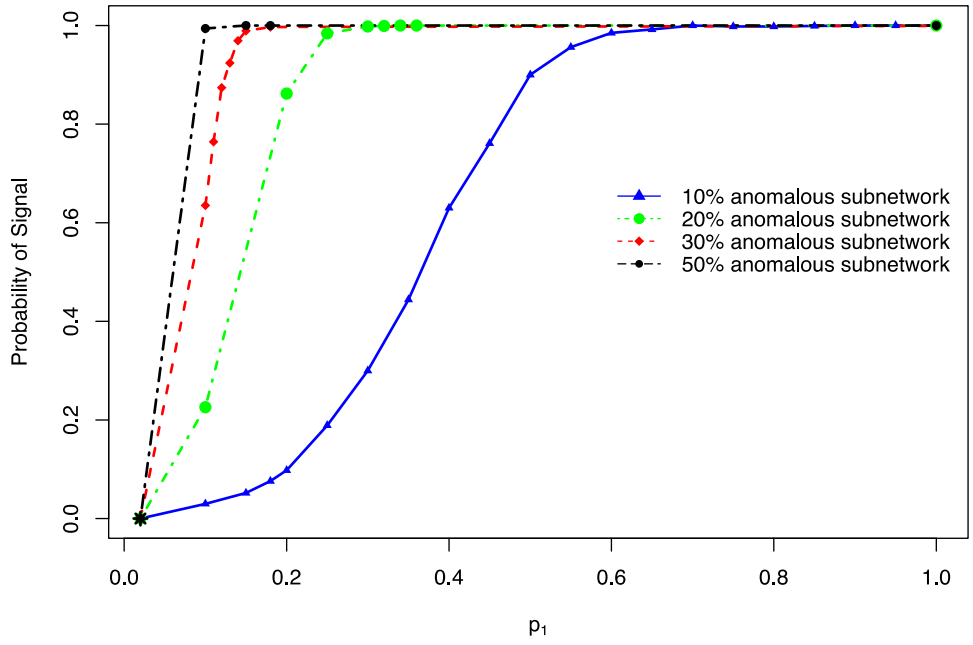

Figure 7. Detection power for a network of order $|V| = 100$ with 10%, 20%, 30%, 50% anomalous sub-networks and with $p_0 = 0.02$ and varying values of p_1. The signal threshold is 5.

Figure 7 shows that when only 10% or less of the social network experiences anomalous behavior, the performance of Priebe's scan method can be poor. A shift from $p_0 = 0.02$ to $p_1 = 0.1$ represents an anomalous event where member nodes are 5 times more likely to contact each other within the anomalous sub-network. In this case, the detection power of the method is only 3.7%. Detection power increases to only 9.1% for the same sub-network when $p_0 = 0.02$ and $p_1 = 0.18$. For an even larger shift to $p_1 = 0.45$, detection power is well below 80%. For smaller anomalous sub-networks, the relative increase from p_0 to p_1 have to be much larger in order to achieve similar SNR values and detection power compared to larger anomalous sub-networks.

In addition, we were also interested to see if a simple method works well. We performed additional simulations that compared Priebe's scan method with simply

monitoring the number of edges within the entire network. Both methods performed well in situations with 50% active sub-networks and the SNR extremely high (> 1000). In realistic situations such as the anomalous sub-network is 10% or smaller the simple method performed poorly regardless of SNR. Priebe's scan method uses the 1st and 2nd neighborhood statistics in conjunction with the degree statistics. Smaller anomalous sub-networks which generate weaker anomaly signals cannot be picked up by only monitoring the degree. The effectiveness of the method is dramatically improved when using all three statistics.

The detection power for the $|V| = 50$ social network is shown in Figure 8. For the 10% anomalous sub-network, detection power is low for all values of p_1, only exceeding a 50 percent detection power for the largest values of p_1. Indeed, even when p_1 = 1, meaning that all members of the clique sub-network are communicating at every time period, there is only about a 60 percent detection power, while the SNR is 12.8. Furthermore, for both 30% and 50% sub-networks, the detection power is 100% when SNR is 20.5, which corresponds to p_1 values of 0.40 and 0.22, respectively.

The detection power for $|V| = 20$ and $|V| = 200$ social networks are shown in the Appendix as Figures 16 and 17. We observed similar trends in detection power for both $|V| = 20$ and $|V| = 200$ social networks. In both cases, when the size of the anomalous sub-network is 10% of $|V|$, the detection power is low and consistent detection of anomalous events by the method is generally not possible. In addition, when comparing similar simulation settings across different social network orders, detection power improves for the larger sub-networks and for very high values of p_1. The detection power of Priebe's scan methods is clearly related to relative size of the anomalous sub-network and the increased level of communication within the sub-network. In realistic situations where the anomalous sub-network is less than 10%

of the network, this method's detection power can be undesirable. The SNR remains relatively small even in the presence of significant shifts from p_0 to p_1 in these scenarios.

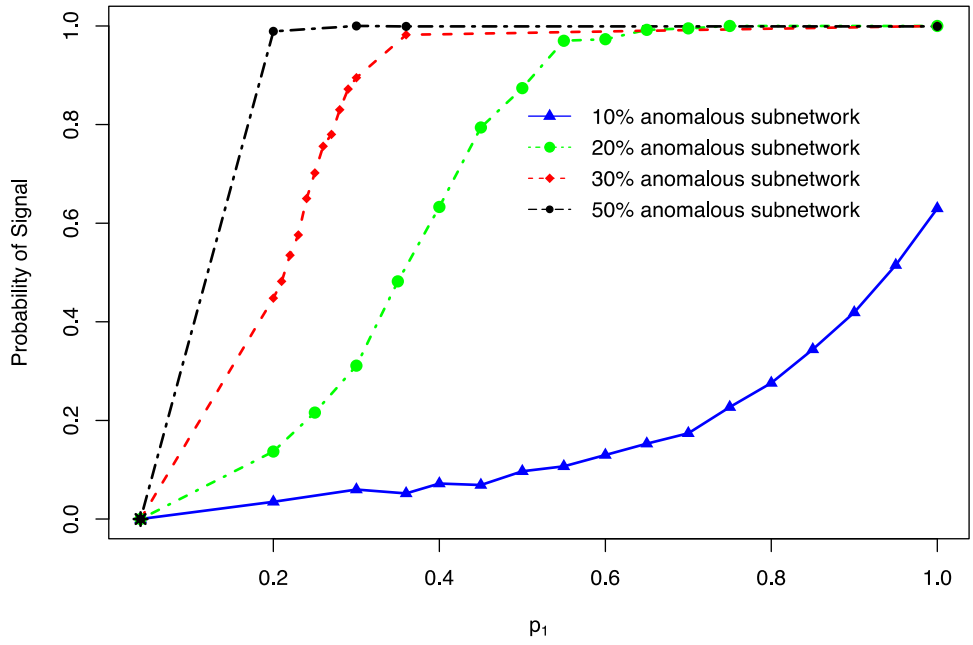

Figure 8. Detection power for a network of order $|V| = 50$ with 10%, 20%, 30%, 50% anomalous sub-networks and $p_0 = 0.1$ and varying values of p_1. The signal threshold is 5.

In general, higher SNR values leads to higher probabilities of detection. Also, similar SNR values yield similar detection power regardless of the anomalous sub-network percentage. Figure 9 shows the relationship between SNR and detection power. Plotted by a selection of the number of anomalous nodes in our simulations. The results shown combine all scenarios with the same number of anomalous nodes, regardless of the simulation settings. For example, network of orders $|V| = 20$ with 50% anomalous sub-network, $|V| = 50$ with 20% anomalous sub-network, as well as $|V| = 100$ with 10% anomalous sub-network all contain 10 anomalous nodes. The relationship between SNR and detection power for all three scenarios are plotted

along a single curve. We observe that the performance of Priebe's scan method is inconsistent across different simulations when $\lambda < 20$. It requires a sufficiently large number of anomalous nodes to work well. Overall, when $\lambda > 20$, Priebe's scan method performs well in our simulations and the detection power approaches 100%. Additionally, we plotted SNR against detection power by the number of anomalous nodes for every simulation scenario in our study. The results are shown in Figure 18 in the Appendix.

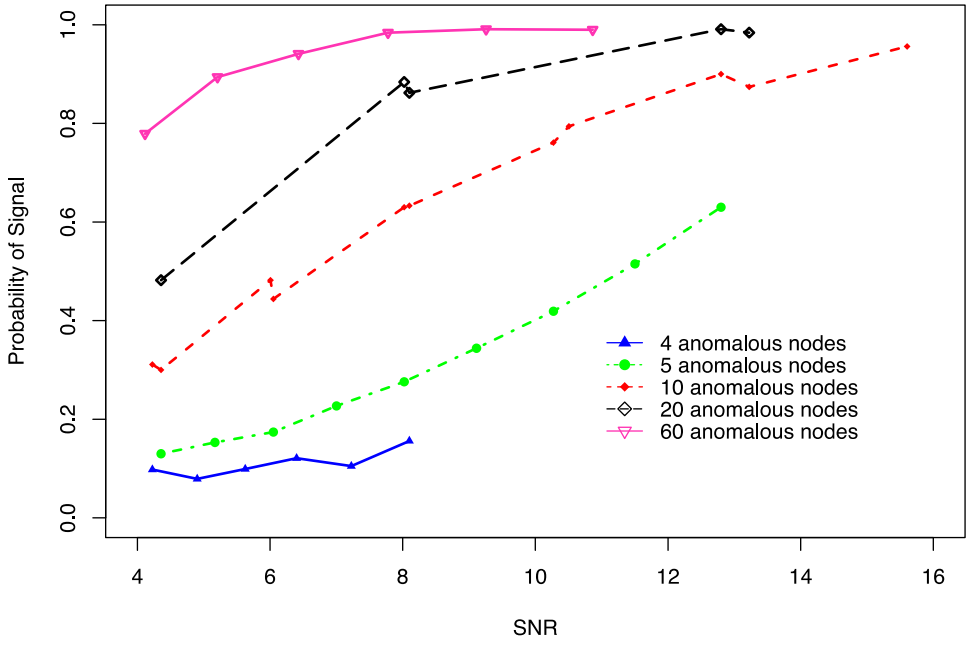

Figure 9. Signal-to-Noise Ratio vs. Detection Power plotted by a selection of the No. of anomalous nodes.

5.3 Moving Window Contamination

The use of a moving window in the scan method means that if the method fails to signal soon after the anomaly, then any anomalous data are quickly incorporated into the baseline. Furthermore, sustained anomalous activity will generally not result in

continuous signaling from the scan method. Thus, as discussed in Section 5.1, signaling becomes increasingly unlikely if anomalies are not detected immediately.

To understand the issue of the contamination of baseline data, we studied the conditional signaling delay of Priebe's scan method. The *conditional signaling delay* is the time from when the anomaly begins ($t = 50$ in our simulations) until the method signals *for those simulations that signaled.* For example, if the method signals at time $t = 51$ then the delay is one time period.

We recorded the delay from 1,000 simulations of a $|V| = 100$ social network, for both 10% and the 30% anomalous sub-networks, with $p_0 = 0.02$ and $p_1 = 0.18$. Figure 10 shows the results, with the 10% sub-network in plot (a) and the 30% sub-network in plot (b), where the histograms show the distribution and the red line the average delay (4.18 time periods for the 10% sub-network and 1.05 time periods for the 30% sub-network).

Figure 10 shows any signal tends to occur immediately after the change in communication rate with this effect more pronounced for the larger anomalous sub-network. It doesn't show, however, that the fraction of times the method correctly signals (i.e., during time $51 \leq t \leq 60$) is also significantly higher for the larger anomalous sub-network. In particular, for the 10% anomalous sub-network shown in plot (a) only 79 of the 1,000 simulations resulted in a signal. The remainder of the simulated methods failed to signal and would have signaled at some time $t > 60$. In contrast, for the 30% anomalous sub-network shown in plot (b) almost all (970) of the 1,000 simulations gave a signal during the time period $51 \leq t \leq 60$. Both the *conditional signaling delay* as well as the probability of detection improve as the SNR value increases.

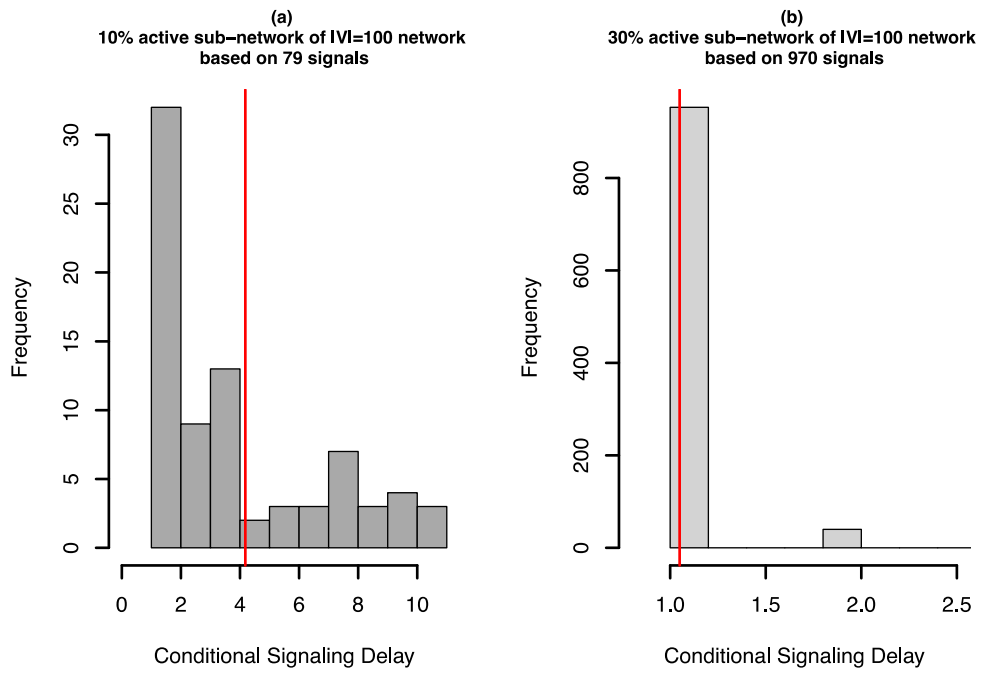

Figure 10. Distributions of conditional signaling delay of 10% and 30% anomalous sub-network in a social network of order $|V| = 100$ with baseline $p_0 = 0.02$ and $p_1 = 0.18$ for t ≥ 51. Signal threshold is 5.

6 Improving Performance

The simulation studies summarized in Section 5 revealed two drawbacks in addition to the method's poor detection power for small anomalous sub-networks. The first is that the method requires extensive amount of baseline data before monitoring can start (in order to establish the moving windows). The second is that any non-signaling anomalous data are quickly incorporated into the moving window, which results in limited opportunities for anomaly detection. However, the following modifications can be made to improve the performance of the method: incorporating a variable sized moving window at the start of monitoring, lowering the signaling threshold, and lagging the moving window.

6.1 Variable Sized Moving Window

Having to collect data for $2w$ time periods in order to establish a baseline for the method is a hindrance when the goal is to start monitoring the network for anomalies as quickly as possible. To mitigate this, we propose using a variable length moving window to lessen the amount of initial data required. This is a commonly used technique in statistical process monitoring with moving average charts, discussed in Montgomery (2009), for combating monitoring delays when the method involves the use of moving windows. The initial number of adjacency matrices needed to establish the moving window can be less than w initially, and gradually increased to w while monitoring progresses.

Using the same simulation settings as Section 5.2, we examined the detection power using a moving window that starts with data from the first 10 time periods. The size of the moving window was then gradually increased to contain the data from 20 time periods, after which the window then subsequently moves with the most recent observation as in Priebe's scan method. The detection power of the $|V| = 100$ order social network obtained using the variable moving window is virtually the same as that indicated in Figure 7, thus illustrating that implementing a variable-width moving window has little detrimental impact on detection power for delayed shifts. A benefit of implementing a variable-width moving window is that it enables the method to detect anomalies during earlier phases of monitoring. Whereas detecting anomalous events in the social network at time $t < 40$ was not possible using Priebe's scan method, the variable window starting at time $t = 10$ allows monitoring to begin 20 time periods earlier.

Since implementing the variable moving window has little detrimental impact other than increasing slightly the number of false alarms, and offers a substantial advantage, we recommend adopting this modification. It will shorten the length of

time prior to the start of monitoring while preserving detection power; monitoring can start following a shorter period of baseline data collection.

6.2 Lowering the Signaling Threshold

We have already established that if the size of the anomalous sub-network is 10% or less, regardless of network order, detection of anomalous events by Priebe's scan method can be poor. Because lowering the threshold increases detection sensitivity, one way to improve detection power is to lower it. In Priebe's scan method, the signaling threshold is fixed at 5. Because the authors provide no justification for using this threshold, we performed simulations to evaluate the influence of the signaling threshold on performance.

A key consideration when choosing a threshold is the inherent trade-off between the sensitivity of the method to detecting actual anomalies and the generation of false positive signals. As the threshold is decreased, sensitivity to detect actual anomalies is improved but that comes at the cost of also increasing the false positive signal rate. Lowering the threshold too far results in an excessive number of false positive signals. Raising the threshold too much results in an insensitivity to detecting actual anomalies (true positive signals).

Figure 11 shows the detection power for a $|V| = 100$ social network with a signaling threshold of 4. For the 10% anomalous sub-network, the improvement in detection power is significant compared to Figure 7 (where the signaling threshold was 5). The detection power improved by nearly 20 percentage points for all p_1 values less than about 0.5. For example, in Figure 11 when $p_1 = 0.4$, the detection power is slightly above 60%. For the same shift with a lower threshold, the detection power is now above 80%. It is important to keep in mind that the SNR values for the networks remain the same regardless of the signaling threshold used.

As such, detection powers are the same for the simulation scenarios with similar SNR values.

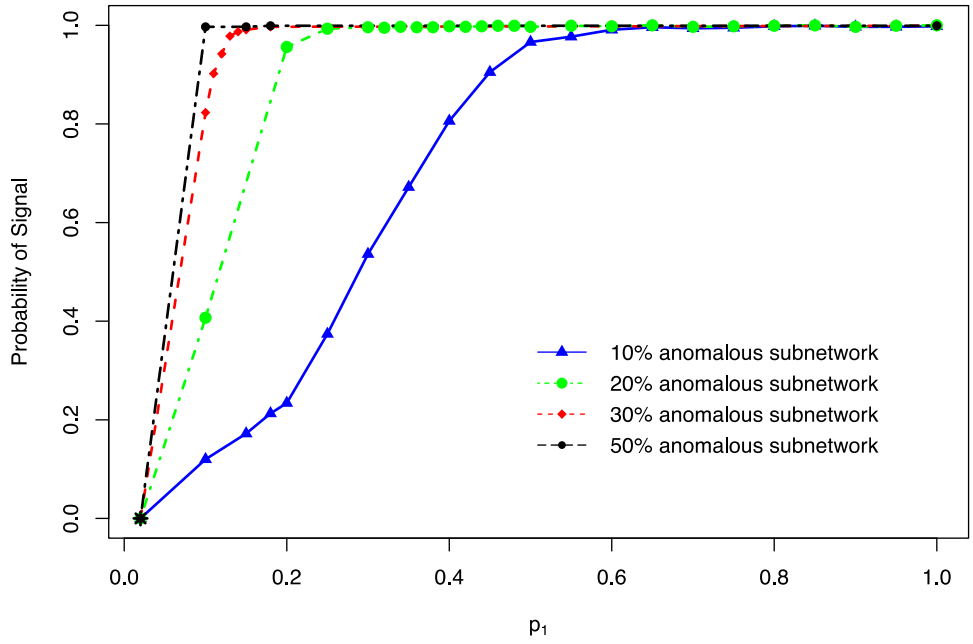

Figure 11. Detection power for a network of order $|V| = 100$ with 10%, 20%, 30%, 50% anomalous sub-network, $p_0 = 0.02$ and varying values of p_1. The signal threshold is set to 4.

Simulations were also run for $|V| = 20, |V| = 50$, and $|V| = 200$ social networks. The detection rates are shown in Figures 19, 20, and 21 respectively in the Appendix. The detection power increased substantially when the threshold was lowered to 4. The results show that there are 5 to 20 percentage points improvements in detection power while using 4 instead of 5 as the signaling threshold. The increase in detection power are observed for all the sub-network percentages studied.

As previously discussed, a consequence of increasing the sensitivity of the method (i.e., lowering the threshold) is that the probability of false signals increases as well. Using signal thresholds of 2, 3, and 4, simulation studies were performed with $|V| = 20, |V| = 50, |V| = 100$, and $|V| = 200$ social networks to assess how

changes in the threshold affect the rate of false signals. The baseline probabilities of communication, p_0, were set to $0.1, 0.04, 0.02,$ and 0.01 to be consistent with the simulation scenarios described in Section 5.2 and, because we are assessing the false positive signaling rate, communication probabilities within the network are kept at baseline levels throughout the simulations. For 1,000 simulations, we recorded the time until the first signal, which was a false signal since no anomaly was present, and then we averaged them to estimate the baseline in-control average time-to-signal (ATOS). Larger baseline in-control ATOS values correspond to lower rates of false positive signals and, conversely, smaller baseline ATOS values correspond to higher rates of false positive signals.

Figure 12 is a plot of the baseline ATOS versus the signal thresholds for $|V| = 20, |V| = 50,\ |V| = 100,$ and $|V| = 200$. In it we see the baseline ATOS is 50 for a signal threshold of 2. It increases to 300 for a threshold of 4, and the projected baseline ATOS using a second order polynomial model for a threshold of 5 is 700, which corresponds to an average of nearly 14 years between false positive signals if the data were aggregated weekly.

Interestingly, Figure 12 shows that the baseline ATOS does not depend on the order of the network, at least in those cases in which p_0 is chosen to keep the expected degree at each node constant and for a constant sub-network percentage. In any case, Figure 12 demonstrates that setting a threshold lower than 5 is reasonable; setting it as low as 3 is perhaps reasonable if one false positive signal per year is tolerable with weekly data. Ultimately, it is up to the practitioner to carefully choose a signal threshold that achieves the highest possible sensitivity for a tolerable false signal rate.

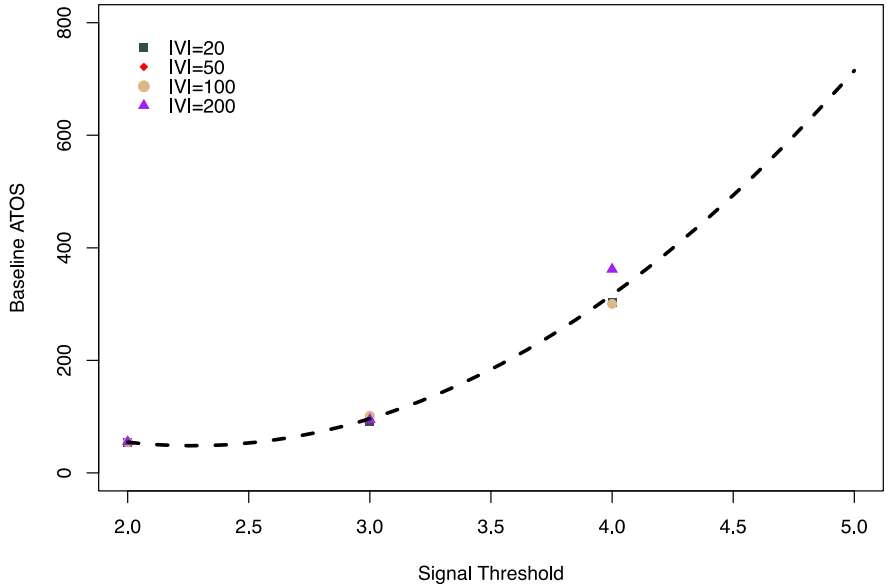

Figure 12. Baseline ATOS vs. Signal Threshold for network orders $|V| = 20, |V| = 50, |V| = 100$ and $|V| = 200$.

6.3 Lagging the Moving Window

With a moving window, the data from any non-signaling period are immediately incorporated into the next moving window. Thus, as previously discussed, when the method fails to detect anomalous events immediately, detection becomes increasingly difficult. We can improve the performance of the method by lagging the moving window, thereby creating additional opportunities for detection. That is, rather than using the previous w time periods for the moving window, where at time t the moving window of data is taken from times $t-1$ to $t-w$, instead incorporate a delay of d time periods, so that the moving window of data is taken from times $t-(1+d)$ to $t-(w+d)$ for some $d = 1, 2, \ldots$.

To assess the effect of adding a lag, we re-ran the simulation of Figure 7 ($|V| = 100$) with a lag of $d = 1$ time period incorporated into the baseline moving

window. Results are displayed in Figure 13. Comparing Figure 13 to Figure 7, immediate improvement in the detection power for all scenarios is evident. For example, the detectability for scenarios with 30% anomalous sub-networks increased by 10 percentage points in overall detection rate when $p_1 \approx 0.1$ or so. When there is a 10% anomalous sub-network with $p_1 = 0.4$ the improvement in detection power is nearly 20 percentage points, nearly 80% instead of 60%. We performed identical simulations for $|V| = 20, |V| = 50$, and $|V| = 200$ social networks as well while implementing the same lag of $d = 1$ time period into the baseline moving window. The detection rates for these simulations are shown in Figures 22, 23, and 24 in the Appendix. We observed similar improvement for these simulations as well. The detection power increased 5 to 10 percentage points for all the simulated social networks.

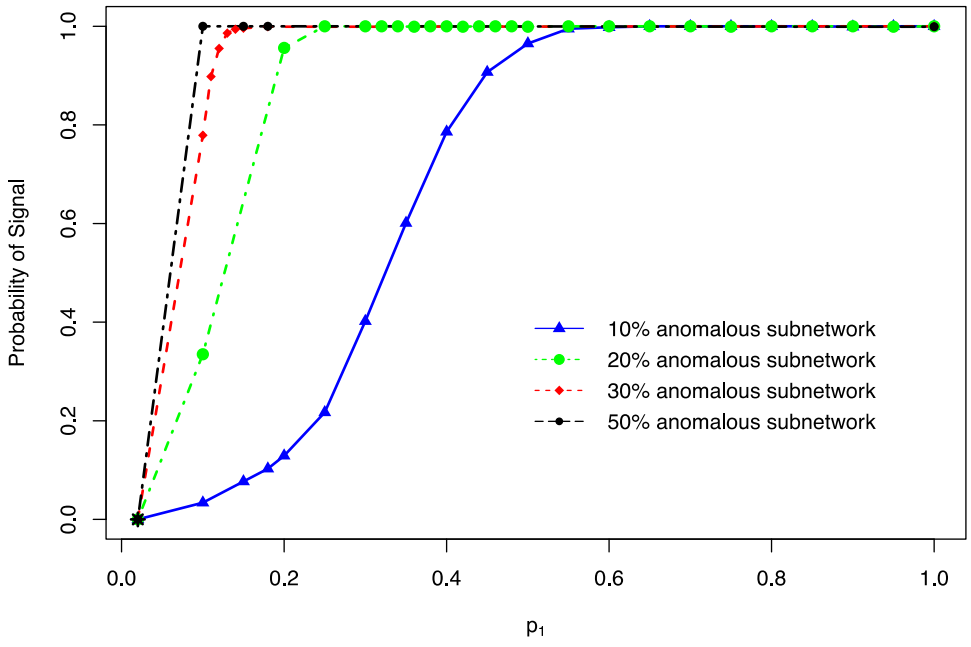

Figure 13. Detectability for a network of order $|V| = 100$ with 10%, 20%, 30%, 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1. Standardization lagged by $d = 1$ time period with the threshold set at 5.

6.4 Evaluation of Improved Scan Method

It is important to note that all three of these improvements can be implemented simultaneously and they will work together synergistically. That is, one can simultaneously modify Priebe's scan method to incorporate a variable sized moving window while appropriately lowering the signaling threshold and lagging the moving window. However, it is also important to recognize that each of the modifications involves a trade-off and they should not be taken to extremes. For the moving window, the beginning window size must be large enough that the sample means and standard deviations can be reasonably estimated or else the standardized values will become excessively noisy, possibly introducing false positive signals. For the selection of thresholds, as discussed in Section 6.2, there is always a trade-off between the sensitivity to true anomalies and the false positive signaling rate. And, for the lagged moving window, for very dynamic networks the greater the lag the less likely that the moving window of data will reflect the correct baseline conditions for the current data. In all three cases, the user must evaluate the particular network being monitored and make reasonable variable window size, threshold, and lag choices.

We evaluated the performance of Priebe's scan method with our suggested improvements, i.e. lowering the signaling threshold, implementing a variable sized moving window with an initial moving window size of 10 and gradually increasing by one to size of 20, and lagging the window. We performed 1,000 simulations for each network scenario listed in Table 1, utilizing a variable sized moving window that lags behind the network data at time t by 1 time point with a signal threshold of 3.

Revising Priebe's scan method to include these three recommendation improves the detection power of the method. Figure 14 shows the detection power for a $|V| = 50$ social network with Preibe's scan method where our suggestions are incorporated compared to the detection power for Preibe's scan method. The original detection power for Preibe's scan method was displayed in Figure 8. There is a 20 percentage point or more increase in detection rate for the 10% anomalous sub-network for all values of p_1. The improvements in detection power are especially notable for smaller shifts in p_1. When $p_1 = 0.2$, detection rate increased 40 percentage points by combining all three improvements for the 30% anomalous sub-network. Comparisons were made for social networks of orders $|V| = 20, |V| = 100$, and $|V| = 200$ were made as well, for which the results are shown in the Appendix as Figures 25, 26, and 27 respectively. We observed similar increases in detection power can be seen for all network scenarios simulated. Overall, compare to the unmodified Priebe's scan method, incorporating our improvement recommendations lowers the SNR values required for achieving similar detection power.

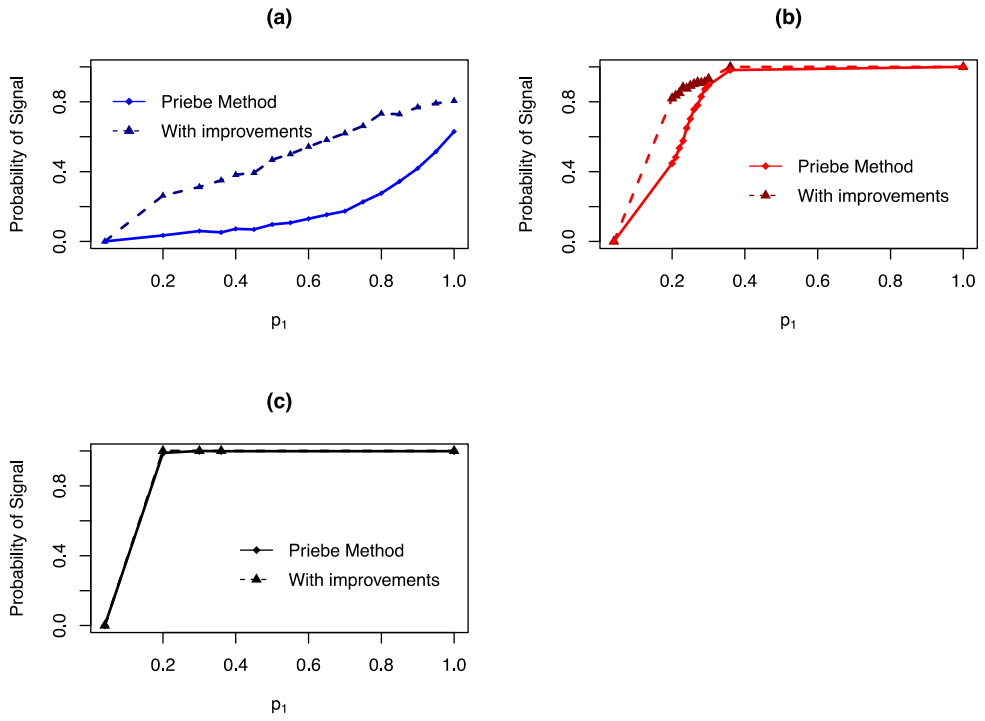

Figure 14. Probability of signal for a network of order $|V| = 50$ with (a) 10%, (b) 30%, (c) 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1. Standardization lagged by $d = 1$ time period, variable moving window size $w = 10$ to 20 with the threshold set at 3.

7 Anomalous Sub-network Identification

While timely detection of anomalous behavior is important in statistical process monitoring, an equally important objective is to be able to identify the nodes associated with this behavior. We have shown that Priebe's scan method with our improvements is able to provide temporal information when anomalies occur, however the method does not provide any sub-network identification.

In this section we propose using a method to identify the sub-network responsible for the anomalous behavior. We evaluate the performance of this method using a simulation study with the same settings described in Table 1. The number of nodes in the networks ranged from 20 to 200 with anomalous sub-network

percentages of $10\%, 20\%, 30\%$ and 50%. The baseline probabilities of communications were scaled based on the number of nodes in the network and ranged $p_0 = 0.1$ to 0.01, and a variety of p_1 values were considered. For each scenario, 1,000 simulations were performed.

The method is initiated following a signal from Priebe's scan method. Once a signal is observed, we determine the time point of the signal and extract the corresponding network snapshot. A clustering algorithm is then implemented to group nodes based on communication patterns among possible sub-networks. For more information on the modularity method we used, see Girvan and Newman (2002). The method identifies the group with the maximum number of communications (i.e., maximum edge count). In practice, this group would be a suggested target for further investigation. We adopted other clustering algorithms to identify the anomalous sub-network as well, however, our results showed that the method by Girvan and Newman (2002) yielded the best results.

In our simulation study, we are able to determine the number and the identities of anomalous nodes in the identified group of interest. Figure 15 displays the true positive rate (TPR) and false positive rates (FPR) using the three values from our simulation of the $|V| = 50$ social network across a range of p_1 values. In cases of 10% anomalous sub-networks, the method is able to capture 60% to 80% of the anomalous nodes. Figure 16 (a) shows TPR increasing from 35% to 60% as the shift probability increases, meanwhile the FPR decreases from less than 20% to near 0%. For 30% anomalous sub-networks, the method captures virtually all anomalous nodes in the social network, translating to 80% TPR and FPR that are nearly 0%. When the anomalous sub-network is 50%, we have 100% TPR and 0% FPR across all shift probabilities. The performance gets better with larger anomalous sub-networks. Similar performance and efficiency of our method is seen in all network scenarios listed in Table 1. The results from these simulations are shown in the

Appendix in Figures 28, 29, and 30. An alternative evaluation method could be based on a metric such as the Dice (1945) similarity metric as used by Megahed et al. (2012) in the monitoring of image data.

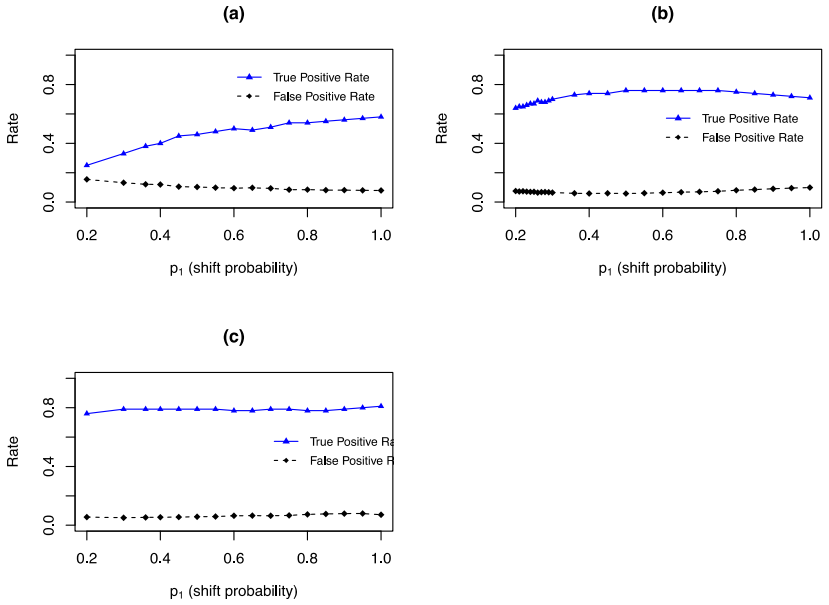

Figure 15. True positive and False positive rates for anomalous sub-network identification for a network of order $|V| = 50$ with (a) 10%, (b) 30%, (c) 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1.

When implemented in conjunction with the Priebe's scan method, Girvan and Newman's (2002) method is able to provide additional information on network anomalies, where we note that its application is not restricted to Priebe's scan method and can be implemented along other network anomaly detection methods as well. Because the majority of current network literature discussions focus on temporal based detection methods, our approach can be a powerful tool when used in conjunction with these methods. Practitioners will have access to both the temporal and sub-network information on network anomalies. This method is

designed to identify possible anomalous nodes post detection, not to improve the performance of the detection method.

8 Conclusions and Future Work

We have evaluated the performance of the moving window based scan method proposed by Priebe et al. (2005), a frequently cited paper in the social network monitoring literature. Using the Erdős–Rényi model, we simulated social networks under various conditions and assessed how well the method signals in the presence of anomalous events. In general, the method can perform poorly when the anomalous sub-network size is a small percentage of the entire network, which is often the type of situation in which a user is most interested. Even when the sub-network is as large as 10% of the entire network the method can have difficulty detecting anomalies. Furthermore, the method is expected to perform even more poorly in realistic monitoring situations where the sub-network size is likely to be much lower than 10%. In particular, the detection power heavily depends upon the signal-to-noise ratio, λ. This parameter measures the relative severity of anomalous events in the network. In scenarios where the SNR value is low, the anomalous signal generated does not represent significant deviations from normal network behavior, or the "noise" in Equation (3). Our simulations have shown that Priebe's scan method is not suitable for detecting anomalous events with low SNR values.

Our evaluation is based on the Erdős–Rényi model, which is especially useful for simulation because it is simple, well known, and easy to grasp intuitively. It may not reflect many of the phenomena present in more realistic networks, however, we have shown that Priebe's scan method does not perform well under certain conditions even in this idealized, simple setting, which raises serious questions about how the method would perform under more realistic conditions and in more

complicated scenarios. More complicated social networks are likely to amplify any shortcomings of the method that have been observed here.

Improvements to Priebe's scan method were proposed, including lowering the signaling threshold, implementing a variable moving window, and lagging the moving window. Our simulations demonstrated that these improvements can significantly improve the method's performance. While Priebe et al. (2005) left unclear whether 5 is a good choice for the signaling threshold, we show that lowering the threshold to 4 significantly improved detection power. A concern for using the lower threshold is that the rate of false signals would increase as a result. However, simulations show that the baseline ATOS using a signaling threshold of 4 can be more than five years when data are aggregated weekly. We recommend that an appropriate signaling threshold should be chosen for improved detection power while taking the baseline ATOS into account. We also demonstrated that adopting a variable moving window to begin monitoring earlier has a negligible effect on the detection rate, where there is no discernable change in the detection power when compared to moving windows of fixed length. The benefit of adopting this modification is that monitoring can start much earlier. For example, detection is possible for anomalies in the social network prior to $t = 40$, while it is impossible to detect such anomalies using the fixed moving window length of $w = 20$ that Priebe et al. (2005) proposed. We also considered lagging the moving window that creates additional opportunity for anomalous data to result in a signal. Simulations with the lagged moving window show that there can be up to a 20 percentage point improvement in detection power for the network scenarios considered. We demonstrated that all three of our improvement recommendations can work synergistically to dramatically improve the performance of the Priebe's scan method.

We recommend using simulated social networks as a general tool with which to evaluate the performance of social network monitoring methods. Using existing

social network data and illustrating performance with a case study for a proposed method is not sufficient. In other process monitoring literature, these types of simulation analyses are accepted and even expected. Furthermore, particularly with anomaly detection methods, the use of simulation is critical because it allows for the control of all aspects of the scenario and, most importantly, the existence, type, and timing of anomalies. As has been discussed, this is a key point because in case study data it is generally not possible to know whether a true anomaly is present or to know precisely when the anomaly began.

Using simulated networks, we recommended a method that can provide anomalous sub-network information post signaling by a temporal based detection method. Our approach can provide the information necessary to identify the anomalous nodes in the network.

Future work will address the limitations inherent in treating network communications within a social network as binary. Clearly there is more information in data that captures the number of communications between network nodes, although there is no reason to believe that the additional count information will overcome the limitations we have identified with Priebe's scan method. In addition, we are planning to assess anomaly detection performance under more realistic models such as the degree corrected stochastic block model (DCSBM), discussed by Karrer and Newman (2011). Wilson et al. (2017) and Yu et al. (2017) have proposed methods for monitoring under the DCSBM.

We agree with Savage et al. (2014) and Woodall et al. (2017) that many comparisons are needed among the various anomaly detection methods that have been proposed. We believe the use of network models is important, where anomalies of interest can be defined in terms of network parameters. Simulation can be used to evaluate the surveillance methods and make comparisons of statistical performance.

Our paper represents a start in this direction using a simple network model and a popular detection method.

References

Akoglu, L., Tong, H., and Koutra, D. (2014). Graph Based Anomaly Detection and Description: A Survey. Data Mining and Knowledge Discovery, 29(3): 626–688.

Azarnoush, B., Paynabar, K., Bekki, J., and Runger, G. C. (2016). Monitoring Temporal Homogeneity in Attributed Network Streams. Journal of Quality Technology, 48(1): 28-43.

Bindu, P. V., and Thilagam, P. S. (2016). Mining Social Networks for Anomalies: Methods and Challenges. Journal of Network and Computer Applications, 68: 213-229.

Carley, K. M. (2003). Dynamic network analysis. Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers, Committee on Human Factors, National Research Council: 133–145

Chapanond, A., Krishnamoorthy M. S., and Yener, B. (2006). Graph Theoretic and Spectral Analysis of Enron Email Data. Computational & Mathematical Organization Theory, 11(3): 265–281.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. Ecology, 26 (3): 297–302.

Erdős, P. and Rényi, A. (1959). On Random Graphs. I. Publicationes Mathematicae, 6: 290–297.

Fricker, R.D., Jr. (2011). Some Methodological Issues in Biosurveillance (with commentaries and rejoinder). Statistics in Medicine, 30: 403-441.

Fricker, R.D., Jr. (2013). Introduction to Statistical Methods for Biosurveillance: With an Emphasis on Syndromic Surveillance. Cambridge University Press.

Fricker, R.D., Jr., Hegler, B.L., and Dunfee D. A. (2008). Comparing Syndromic Surveillance Detection Methods: EARS' Versus a CUSUM-Based Methodology. Statistics in Medicine, 27: 3407-3429.

Fricker Jr., R. D., Knitt, M. C., and Hu, C. X. (2008). Comparing Directionally Sensitive MCUSUM and MEWMA Procedures With Application to Biosurveillance. Quality Engineering, 20(4): 478-494.

Girvan, M. and Newman, M. E. J. (2002). Community Structure in Social and Biological Networks. Proceedings of the National Academy of Sciences, 99(12): 7821-7826.

Hajek, B., Wu, Y., and Xu, J. (2015). Recovering a Hidden Community Beyond the Spectral Limit in O(|E|log*|V|) Time. *arXiv preprint arXiv:1510.02786*.

Hajek, B., Wu, Y., and Xu, J. (2017). Information limits for recovering a hidden community. IEEE Transactions on Information Theory. 63(8): 4729-4745.

Karrer, B. and Newman, M. E. J. (2011). Stochastic Blockmodels and Community Structure in Networks. Physical Review E, 83(1): 016107.

McCulloh, I., and Carley, K. M. (2011). Detecting Change in Longitudinal Social Networks. Journal of Social Structure, 12(3).

Megahed, F. M., Wells, L. J., Camelio, J. A., and Woodall, W. H. (2012). A Spatiotemporal Method for the Monitoring of Image Data. Quality and Reliability Engineering International, 28(8): 967-980.

Montanari, A. (2015). Finding One Community in a Sparse Graph. Journal of Statistical Physics, 161(2): 273-299.

Montgomery, D. C. (2012). Introduction to Statistical Quality Control. 7th Ed., John Wiley & Sons: 442-444.

Mousavi, S. and Reynolds Jr, M. R. (2009). A CUSUM Chart for Monitoring A Proportion With Autocorrelated Binary Observations. Journal of Quality Technology, 41(4): 401-414.

Priebe, C. E., Conroy, J. M., Marchette, D. J., and Park, Y. (2005). Scan Statistics on Enron Graphs. Computational & Mathematical Organization Theory, 11(3): 229-247.

Ranshous, S., Shen, S., Koutra, D., Harenberg, S., Faloutsos, C., and Samatova, N. F. (2015). Anomaly detection in dynamic networks: a survey. Wiley Interdisciplinary Reviews: Computational Statistics, 7(3): 223-247.

Savage, D., Zhang, X., Yu, X., Chou, P., and Wang, Q. (2014). Anomaly Detection in Online Social Networks. Social Networks, 39: 62-70.

Wei, W., and Carley, K. M. (2015). Measuring Temporal Patterns in Dynamic Social Networks. ACM Transactions on Knowledge Discovery from Data (TKDD), 10(1): 1-27.

Wilson, J.D., Stevens, N.T., and Woodall, W.H. (2017). Modeling and Detecting Change in Temporal Networks via a Dynamic Degree Corrected Stochastic Block Model. Submitted for publication.

Woodall, W. H., Zhao, M. J., Paynabar, K., Sparks, R., and Wilson, J. D. (2017). An Overview and Perspective on Social Network Monitoring. IISE Transactions, 1-12.

Yu, L, Woodall, W. H., and Tsui, K. L. (2017). "Detecting Node Propensity Changes in the Degree-corrected Stochastic Block Model", Submitted for publication.

Appendix

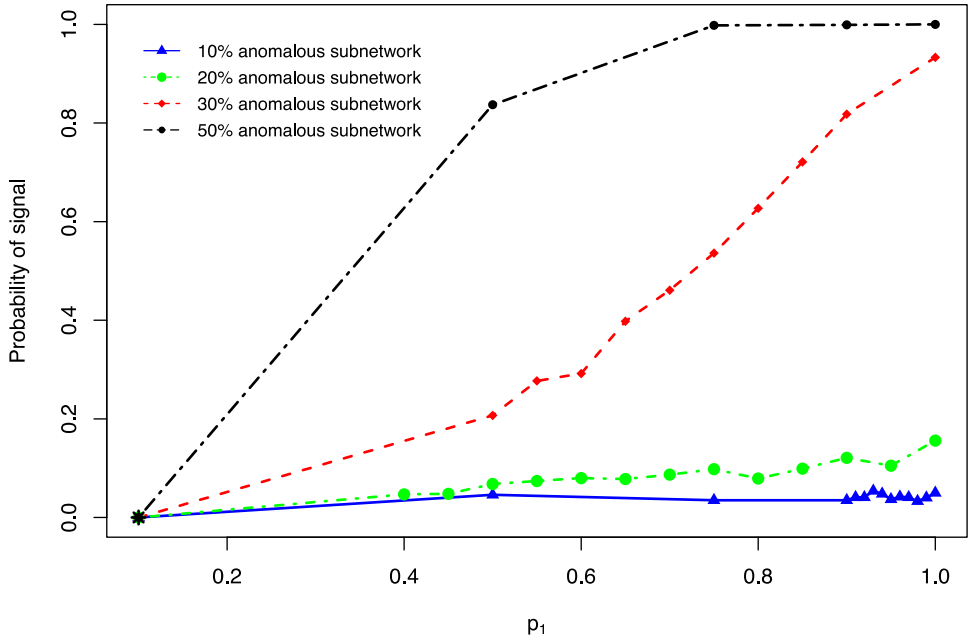

Figure 16. Detection power for a network of order $|V| = 20$ with 10%, 20%, 30%, 50% anomalous sub-networks and with $p_0 = 0.02$ and varying values of p_1. The signal threshold is 5.

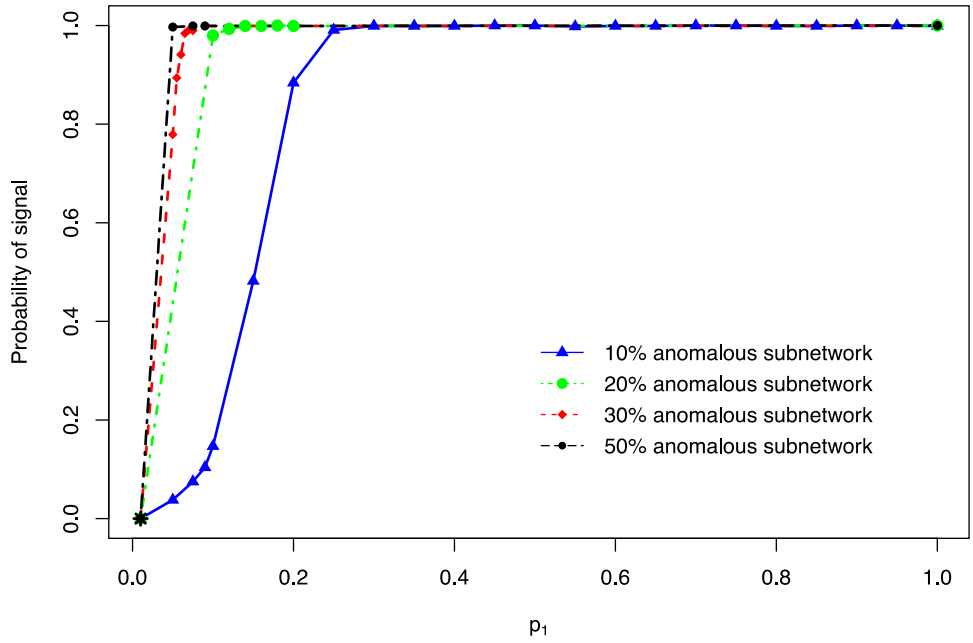

Figure 17. Detection power for a network of order $|V| = 200$ with 10%, 20%, 30%, 50% anomalous sub-networks and with $p_0 = 0.02$ and varying values of p_1. The signal threshold is 5.

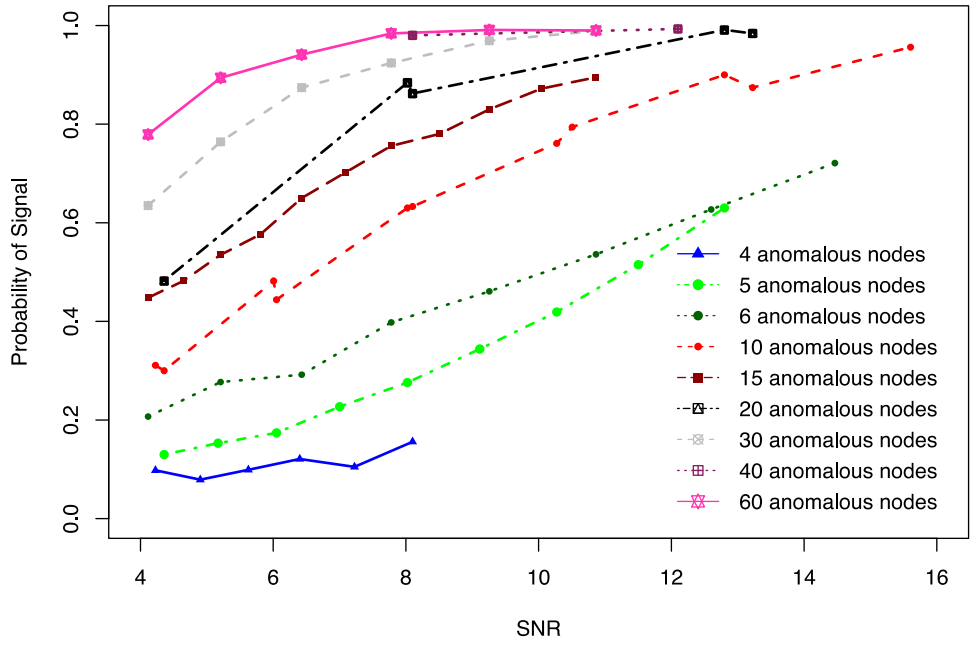

Figure 18. Signal-to-Noise Ratio vs. Detection Power plotted by the No. of anomalous nodes in all simulation scenarios.

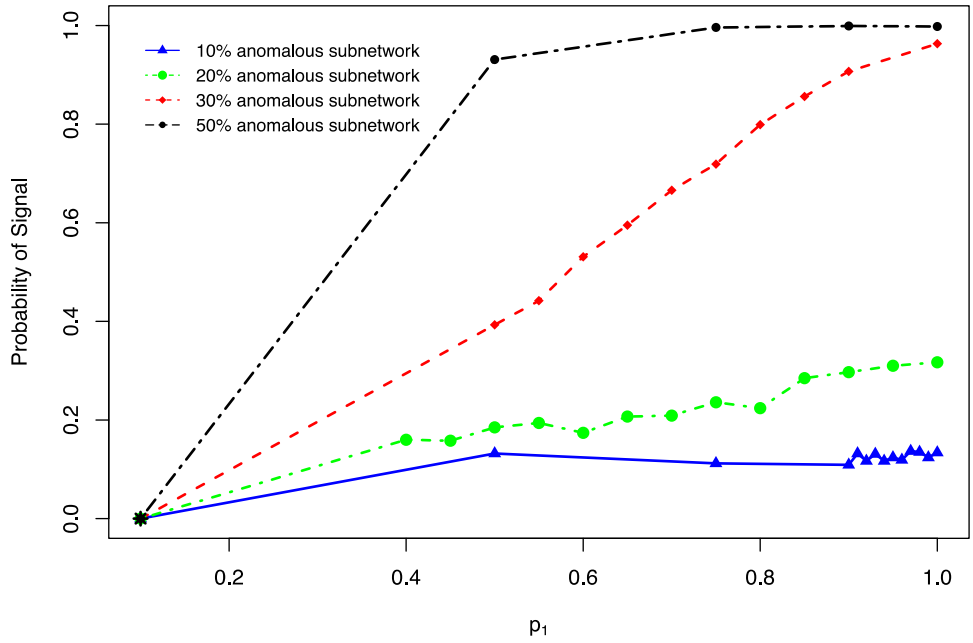

Figure 19. Detection power for a network of order $|V| = 20$ with 10%, 20%, 30%, 50% anomalous sub-network, $p_0 = 0.02$ and varying values of p_1. The signal threshold is set to 4.

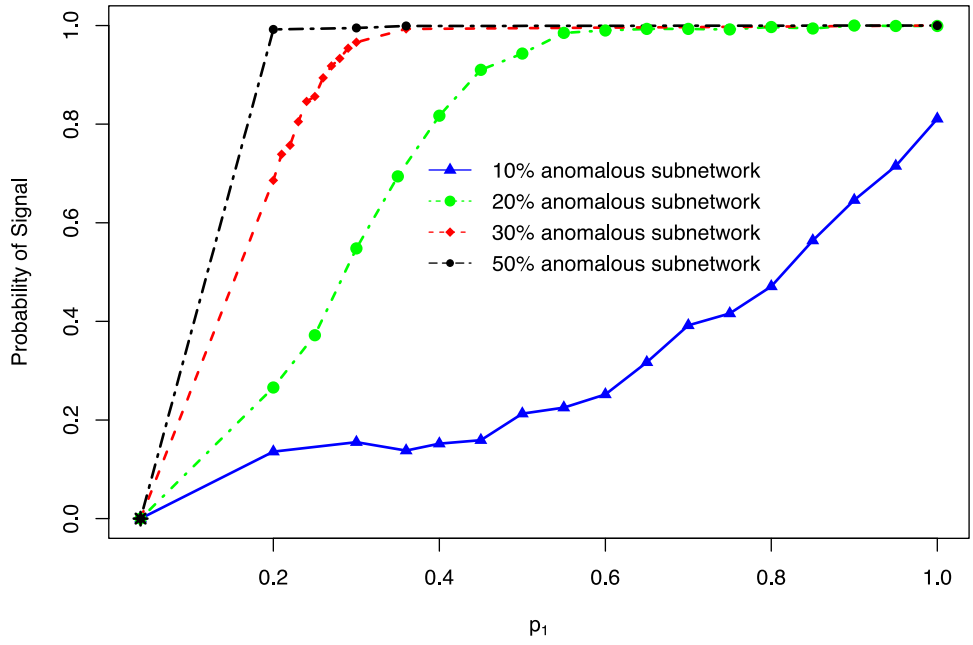

Figure 20. Detection power for a network of order $|V| = 50$ with 10%, 20%, 30%, 50% anomalous sub-network, $p_0 = 0.02$ and varying values of p_1. The signal threshold is set to 4.

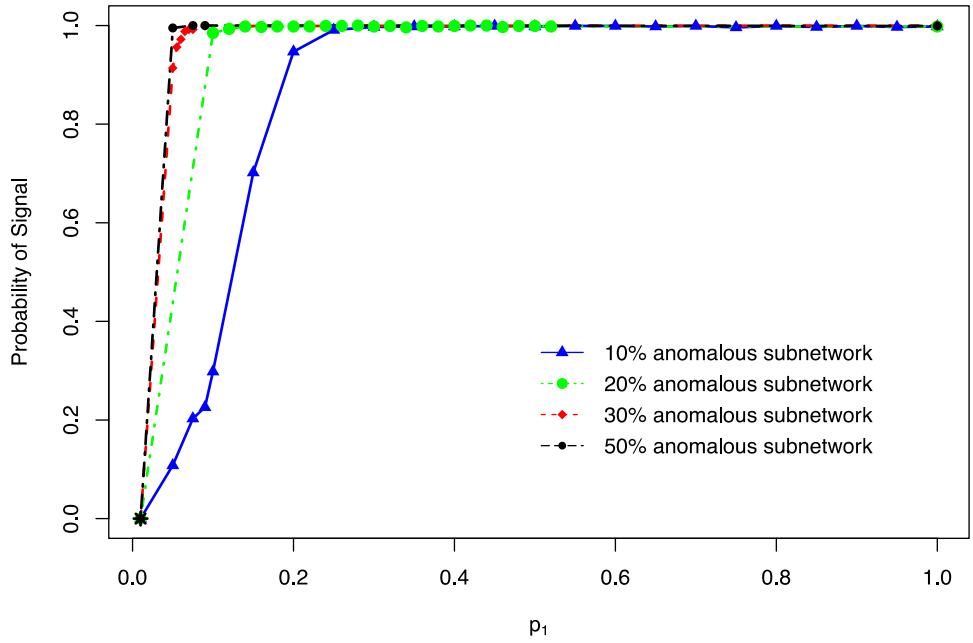

Figure 21. Detection power for a network of order $|V| = 200$ with 10%, 20%, 30%, 50% anomalous sub-network, $p_0 = 0.02$ and varying values of p_1. The signal threshold is set to 4.

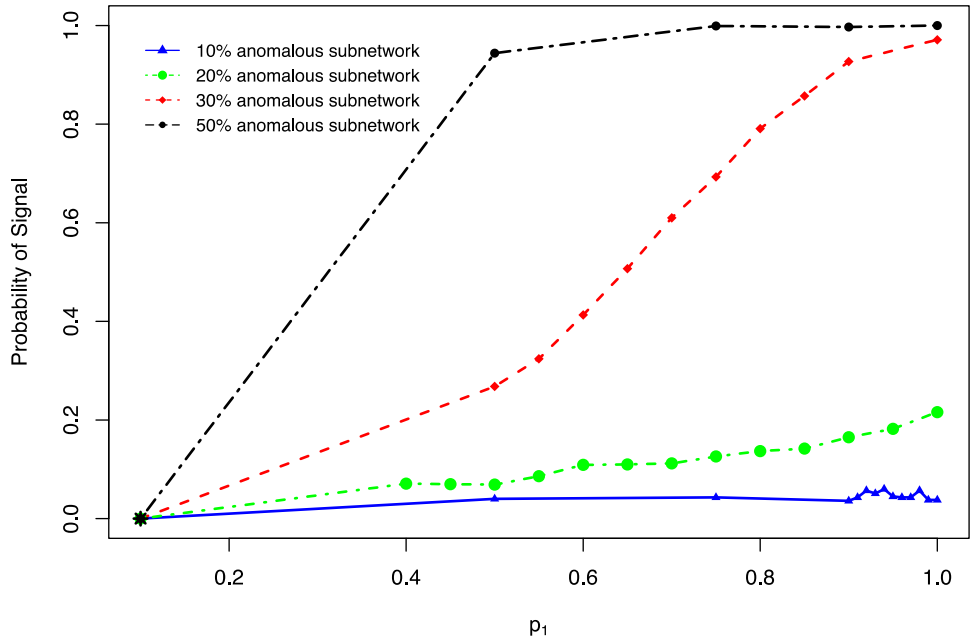

Figure 22. Detectability for a network of order $|V| = 20$ with 10%, 20%, 30%, 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1. Standardization lagged by $d = 1$ time period with the threshold set at 5.

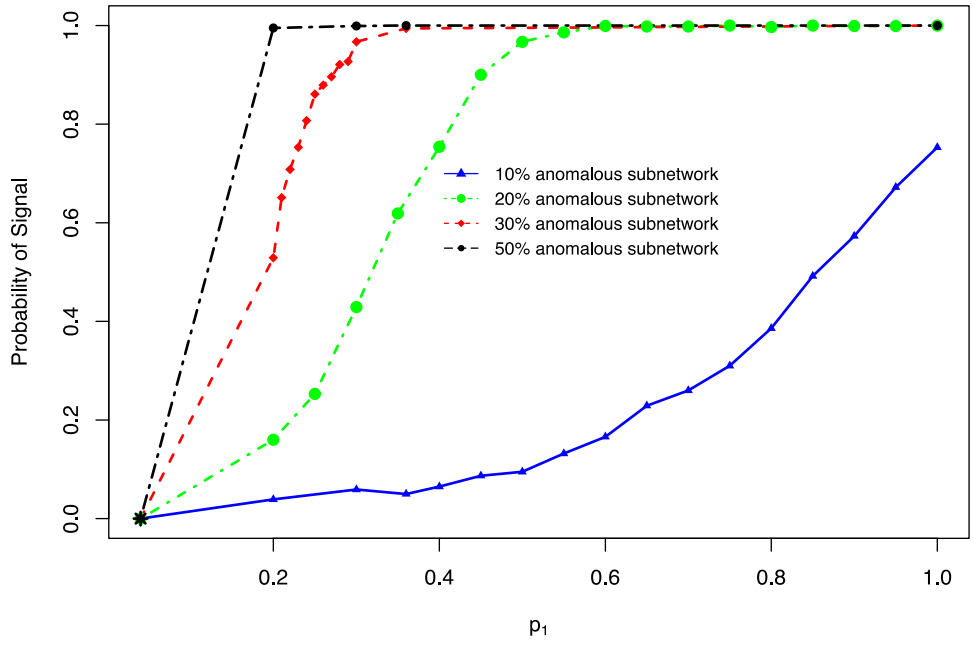

Figure 23. Detectability for a network of order $|V| = 50$ with 10%, 20%, 30%, 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1. Standardization lagged by $d = 1$ time period with the threshold set at 5.

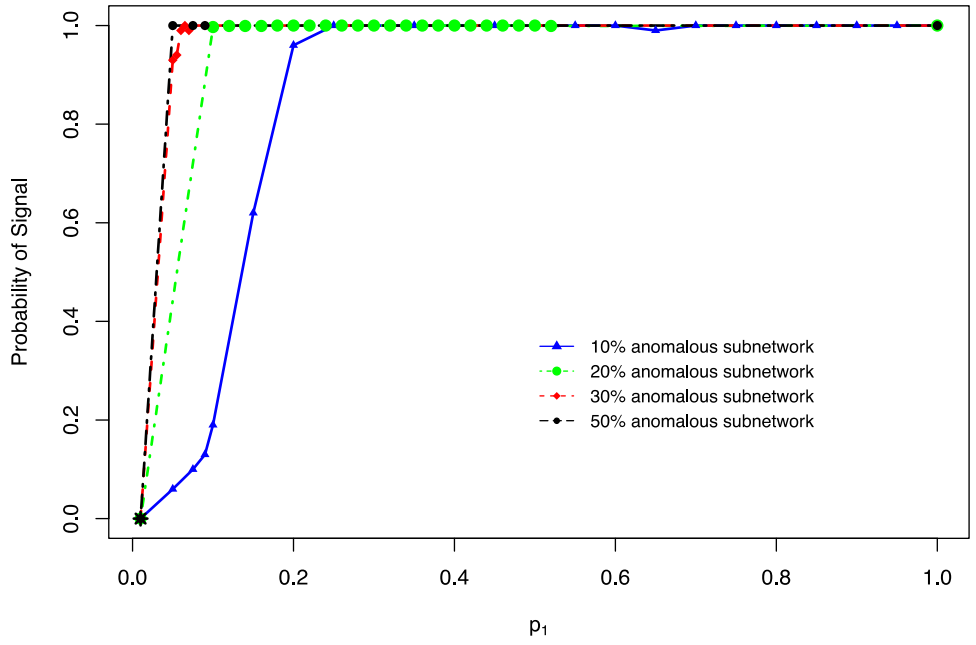

Figure 24. Detectability for a network of order $|V| = 200$ with 10%, 20%, 30%, 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1. Standardization lagged by $d = 1$ time period with the threshold set at 5.

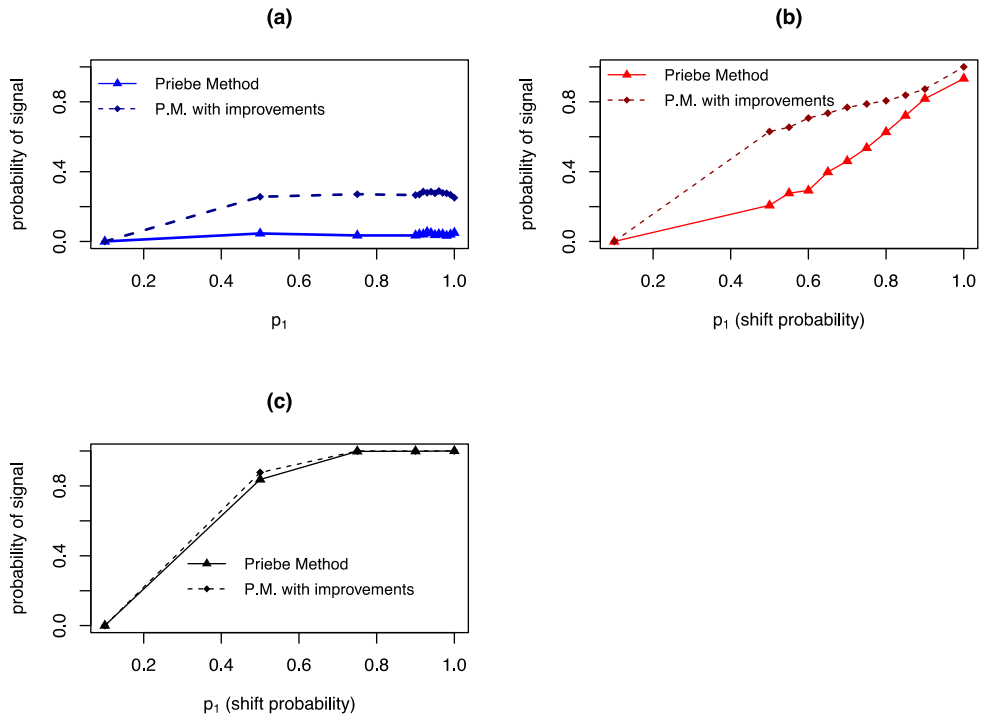

Figure 25. Probability of signal for a network of order $|V| = 20$ with (a) 10%, (b) 30%, (c) 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1. Standardization lagged by $d = 1$ time period, variable moving window size $w = 10$ to 20 with the threshold set at 3.

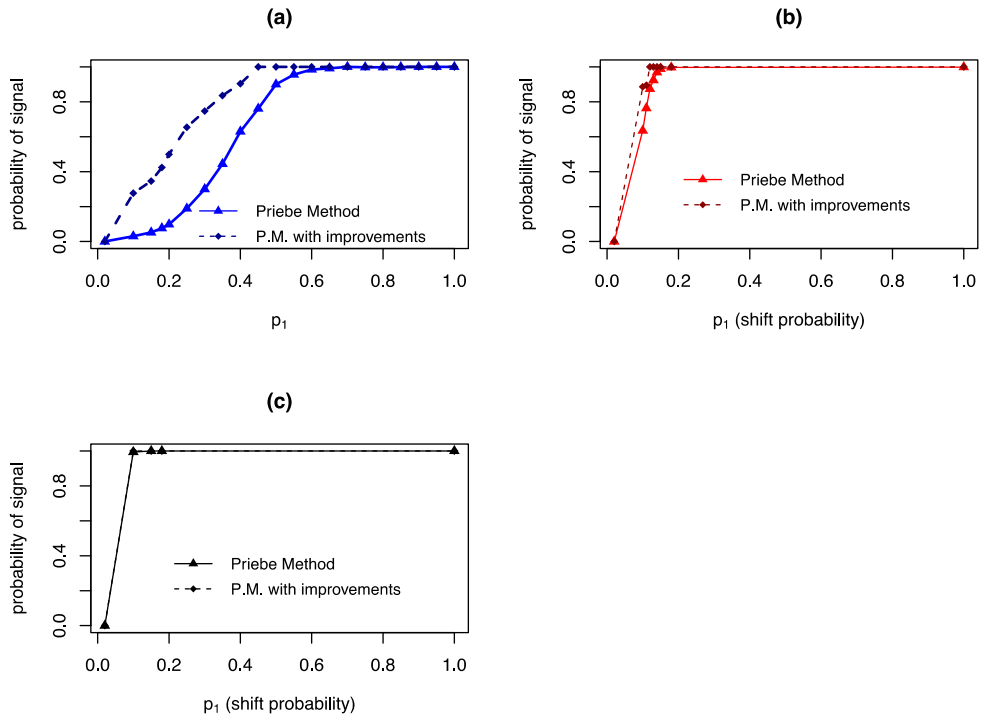

Figure 26. Probability of signal for a network of order $|V| = 100$ with (a) 10%, (b) 30%, (c) 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1. Standardization lagged by $d = 1$ time period, variable moving window size $w = 10$ to 20 with the threshold set at 3.

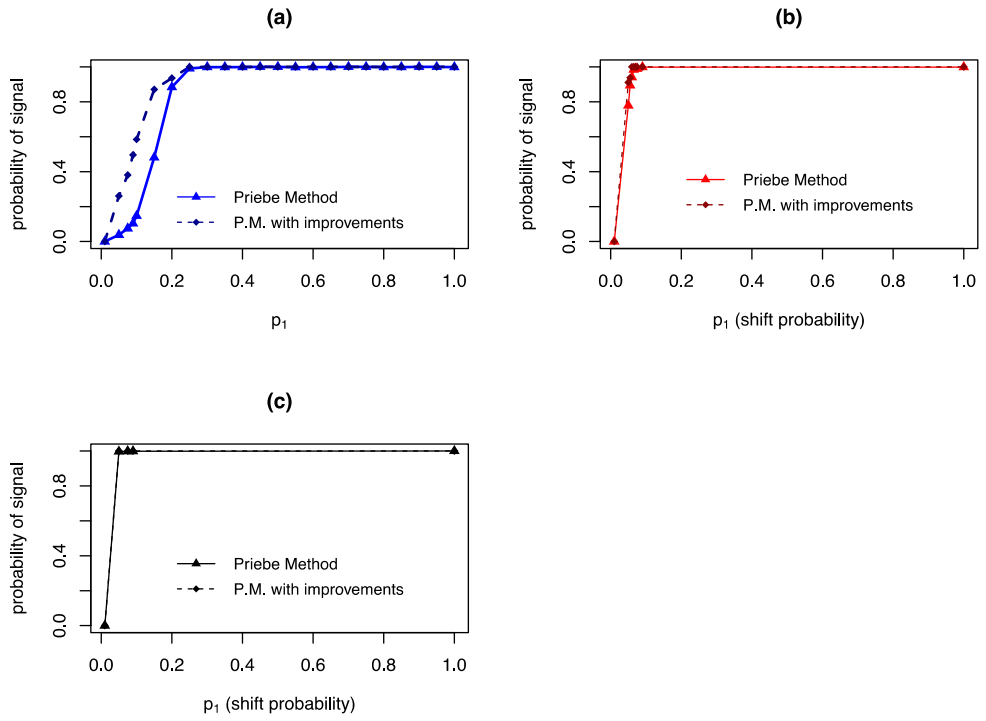

Figure 27. Probability of signal for a network of order $|V| = 200$ with (a) 10%, (b) 30%, (c) 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1. Standardization lagged by $d = 1$ time period, variable moving window size $w = 10$ to 20 with the threshold set at 3.

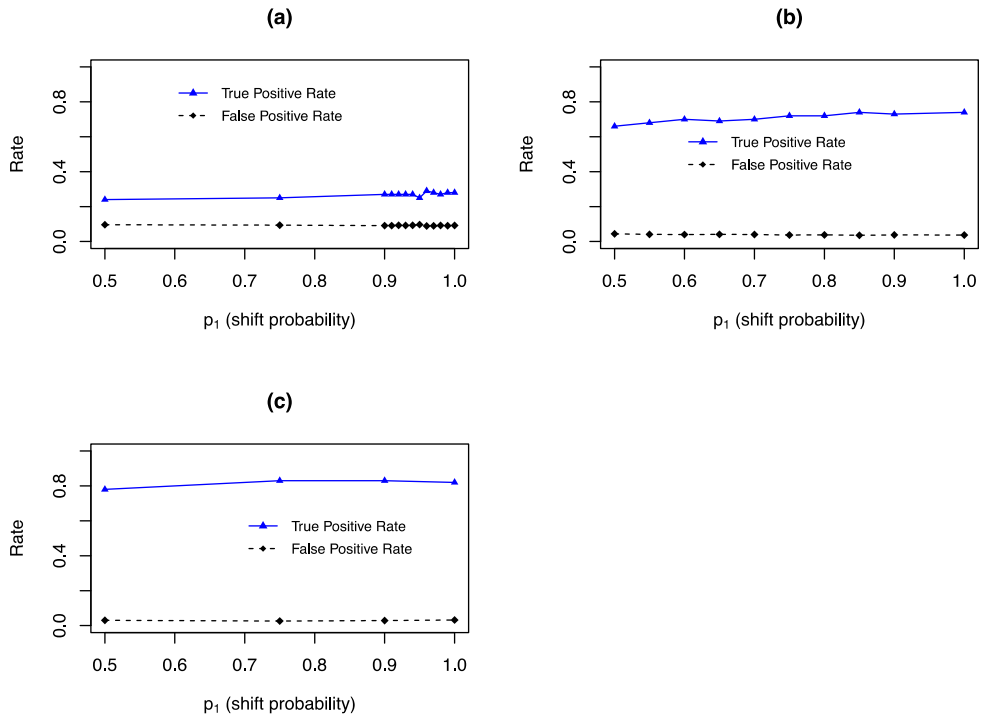

Figure 28. True positive and False positive rates for anomalous sub-network identification for a network of order $|V| = 20$ with (a) 10%, (b) 30%, (c) 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1.

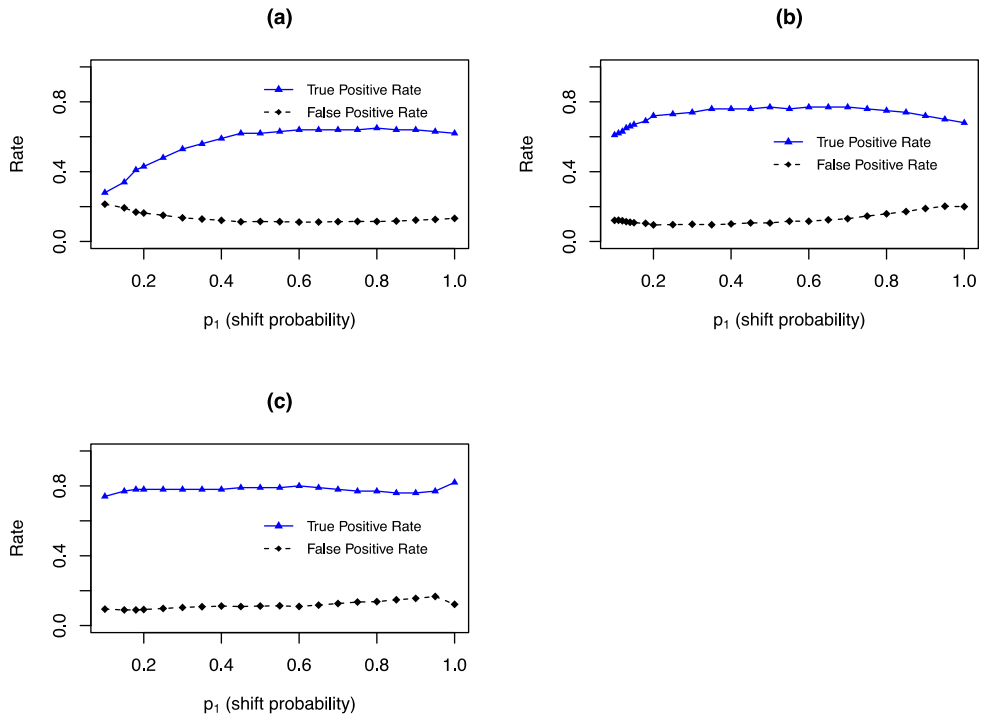

Figure 29. True positive and False positive rates for anomalous sub-network identification for a network of order $|V| = 100$ with (a) 10%, (b) 30%, (c) 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1.

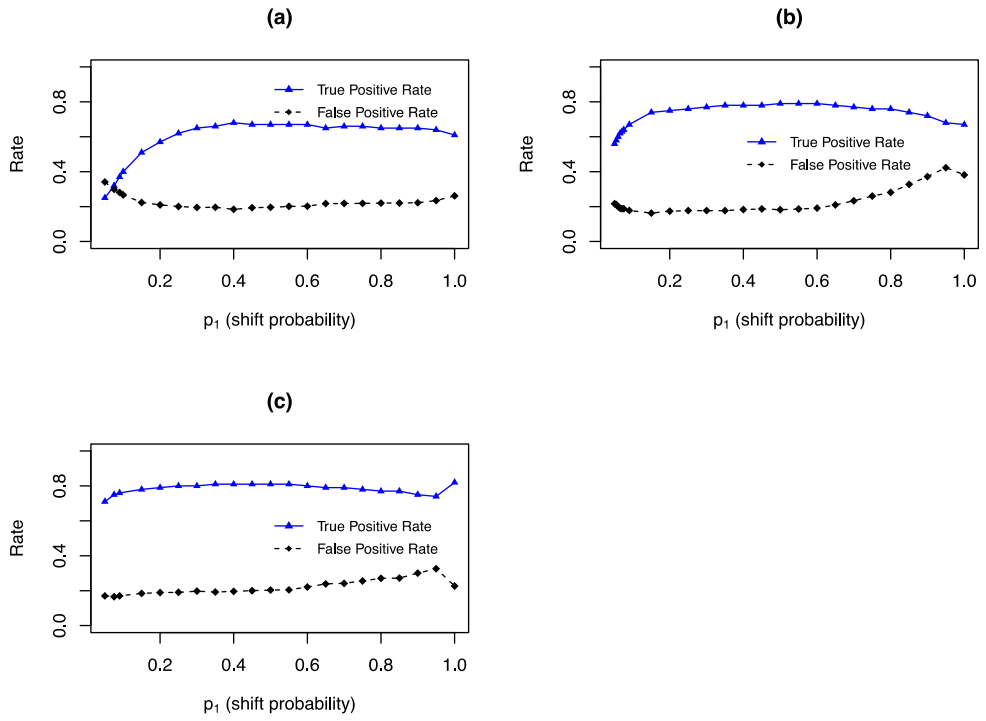

Figure 30. True positive and False positive rates for anomalous sub-network identification for a network of order $|V| = 200$ with (a) 10%, (b) 30%, (c) 50% anomalous sub-networks and $p_0 = 0.02$ and varying values of p_1.

Chapter 3

Data Aggregation in Social Networks

In Chapter 2, we evaluated the performance of the Priebe's scan method using simulated network data and provided our recommendations for improvement to the method. Additionally, we studied a modularity method to identify the anomalous nodes in the network. In Chapter 4, we continue our investigation of social network anomaly detection by researching an important issue, data aggregation. Non-aggregated network data exist only in situations where the precise times for all communication instances in the network are known. For network surveillance methods such as Priebe's scan method, however, the network data are always aggregated. We study how different aggregation levels affect the performance of Priebe's scan method. Through our investigation, we provide valuable insights to practitioners on how the detection method is affected based on the aggregation level of the network data.

Chapter 4

The Effect of Data Aggregation Level in Social Network Monitoring

Meng J. Zhao, Anne R. Driscoll, Srijan Sengupta, Nathaniel T. Stevens, Ronald D. Fricker Jr., William H. Woodall (2017). Submitted for publication.

Abstract

Periodic data collection is an essential part of any real time monitoring process. In doing so, the data collected are almost always aggregated at a specific level. There are numerous studies on aggregated data for monitoring events of interest in the process monitoring literature, however, the effect of data aggregation in social network surveillance has not been thoroughly investigated. In this paper, we employ the Degree Corrected Stochastic Block Model (DCSBM) to simulate social networks, and network anomalies based on the intrinsic community structure within the network. Network data simulated using DCSBM are presented in both count and binary formats. In conjunction, we use Priebe's scan method as the monitoring method. Our research show that when the availability of post-anomaly data is limited and the data is aggregated at high levels, there is considerable decrease in anomaly detection performance. Moreover, there is significant loss of information that is detrimental to network anomaly detection if the social network communication data are in binary form. In contrast, for network count data, aggregation at an appropriate level can amplify the anomalous signal generated by the network anomalies and increase detection performance. While not comprehensive, our results and recommendation can be generalized to other situations, and serve as a guideline for practitioners in network surveillance situations that involve aggregated network data.

Keywords aggregation, DCSBM, scan method, social network change detection, statistical process monitoring

1 Introduction

The goal of social network surveillance is to be able to detect and locate anomalous events in the network in a timely fashion. Social network data consist of counts of communications between individuals or indictor variables indicating the presence of absence of a given level of communication aggregated over a given time interval such as day or week. The aggregation interval length affects the amount of information that is contained within the reported data. Therefore, the level of aggregation impacts the performance of any monitoring technique. Zhao et al. (2017) evaluated a popular moving window based scan method introduced by Priebe et al. (2005) at a fixed level of aggregation. In this paper, we study the effect of data aggregation and the effect of conversion of count data to binary data on social network surveillance using this method as an illustration.

There are several studies on data aggregation for monitoring events of interest in the process monitoring literature. Benneyan (1998) discussed data aggregation in healthcare surveillance and recommended aggregating data over short time periods with an emphasis on collecting data as often as possible. Schuh et al. (2014) studied the effect of aggregating count data under a Poisson process model. Dubrawski and Zhang (2010) discussed the practical use of data aggregation when public health data availability is limited. However, the effect of data aggregation in social network surveillance applications has not been investigated.

Social network surveillance involves two types of network data, binary and count. Network count data in the form of $\{0, 1, 2 \ldots\}$ indicates the number of communication instances between pairs of individuals. Network binary data in the form of $\{0, 1\}$ is used to indicate when the number of communications between pairs of individuals in the social network surpass a certain threshold. This threshold can simply indicate the presence of any communication, or any other appropriate level

where 0 represents communication level less than the threshold, and 1 represents communication level equal to or more than the threshold.

When reviewing monitoring methods for detecting events of interest using temporal data, Woodall et al. (2008) stated that aggregating data over time results in a loss of information and slower detection. Our research shows that this information loss is greater if the social network communication data are in binary form, however the problem is not as severe when communication data are recorded as count data. Aggregating count data involves combining the communication counts cumulated over the aggregation period. Information on the difference in communication counts among pairs of individuals over time are retained. As continuous data collection and analysis would be more difficult, network surveillance almost always involves aggregated data. While not comprehensive, our investigation includes a number of important factors which we think represent some common scenarios that practitioners are likely to encounter. We will show that aggregation in certain situations can amplify the anomalous events in the social network, increasing the detection sensitivity of the method used. Furthermore, we show that too much aggregation can lower the performance, and that the delay in signaling caused by aggregation can be disadvantageous.

Different types of statistical models can be used to simulate social network data. The Erdős–Rényi (Erdős and Rényi 1959) model used by Zhao et. al. (2017) is a simple model, and provides an excellent starting point for studying social networks using simulation. Nevertheless, this model ignores two important and natural aspects of social networks, community structure and communication variability among the individuals in the network. Communities partition the individuals in a social network such that the propensity for communication between individuals within a community is larger than between communities. The degree corrected stochastic block model (DCSBM) was introduced by Karrer and Newman (2011) and used by Wilson et. al.

(2017) for network surveillance applications. The DCSBM takes both community structure and individual communication behavior into account. Generally, the DCSBM is better suited for simulating realistic social networks than the Erdős–Rényi (Erdős and Rényi 1959) model. In our investigation, we employ the DCSBM in conjunction with the Priebe et al. (2005) method to study the effect of aggregation on social network surveillance. We draw general conclusions that can be applied to other network surveillance strategies as well.

The remainder of this paper is organized as follows. In Section 2, we introduce common network terminology. This is followed in Section 3 by a description of the DSCBM model used to simulate social network data, as well as the Priebe's scan method used to monitor the networks. We introduce the simulation scenarios studied in Section 4, and examine the false alarm rates resulting from data aggregated at different levels in Section 5. Simulation results for network data are discussed in Section 6. In Section 7, we study the conditional signal delays caused by aggregation, and draw our conclusions in Section 8.

2 Network Terminology

We will define social network terminology in this section for the general reader who may be unfamiliar with both the terminology and notation. A social network $G = (V, E)$ consists of a set of *nodes* V and a set of *edges* E. In a social network, a node usually represents an individual or the origin/destination of a communication. When the nodes in a social network contain information that makes them readily identifiable, such as a name or an e-mail address, the social network is said to be *labeled*. If no identifying information is available, it is an *unlabeled* social network. Depending on the type of network data, the edges in the social network specify the level of communication between each pair of nodes over a given period of time. For network count data, the edges represent the number of communication instances

between each pairs of nodes. For network binary data, each edge represents an indication that communication between a pair of nodes exceeded a certain threshold. We denote the number of nodes and edges by $|V|$ and $|E|$, respectively. The number of nodes, $|V|$, is referred to as the *order* of the network. The number of edges, $|E|$, is referred to as the *size* of the network.

The communication levels between all pairs of nodes at time t for a network $G_t = (V, E_t)$ is summarized by an adjacency matrix C_t of dimension $|V| \times |V|$. The matrices C_t are available at the current time ($t = T$) and all previous time points, $t = 1, 2, \ldots, \mathrm{T} - 1$, each representing a snapshot of the social network structure at these specific times. The entry of C_t in row i and column j, denoted by $C_t(i, j)$, is a record of the communication level from node i to node j. Thus, the direction of communication is indicated by the quadrant, upper right or lower left, of off-diagonal entries where $C_t(i, j)$ lies. Networks for which the direction of communications are known are referred to as directed networks. Values on the diagonal are set to zero. When the direction of communication is not measured or unknown we have $C_t(i, j) = C_t(j, i)$ and C_t is a symmetric matrix. Networks in case are referred to as undirected networks. We will specifically focus on undirected networks in our investigation. When only the presence or absence of a specified level of communication is measured, each $C_t(i, j)$ is restricted to the values 1 (presence) or 0 (absence). When the total number of communications is measured, each $C_t(i, j)$ can take any non-negative integer value.

The k^{th} *neighborhood* of node i contains all nodes and edges that are within k edges of node $i, k = 1,2,3, \ldots$. The *size* of the k^{th} neighborhood of node i is the count of all edges between nodes in the k^{th} neighborhood of node i. At time t, this quantity is denoted by $O_{t,i}^{k}$. We use the quantity $O_{t,i}^{0}$ to represent the *degree* of node i, i.e., the number of communications to or from node i at time t.

3 Model and Method

3.1 Degree Corrected Stochastic Block Model

The DCSBM provides better insight into how data aggregation affects anomaly detection than the Erdős–Rényi model for two reasons. First, anomalous events in social networks are typically associated with changes in communication patterns within and/or between communities. Second, individuals throughout the network tend to be heterogeneous and have varying propensities for communication. Since the DCSBM takes both community structure and node degree heterogeneity into account, it is a useful model for simulating realistic social networks to study the effects of data aggregation.

Under the DCSBM model, the $|V|$ nodes in a social network are partitioned into R disjoint communities where community, $r = 1, 2, \dots, R$ contains $|V_r|$ nodes and $V = V_1 \cup V_2 \cup \cdots \cup V_R$. The following parameters are specified by the DCSBM model: the parameter θ_i, $\{i = 1, \dots, |V|\}$ defines the propensity of node i to communicate with any other node, an $R \times R$ matrix P defines the propensity of communication among the different communities over a given period of time, where element $P_{r,r'}, \{r, r' = 1, \dots, R\}$ is the communication propensity between communities r and r'. The number of communications between nodes i and j is then assumed to be Poisson distributed with mean

$$\lambda_{i,j} = \theta_i \theta_j P_{r,r'}\,, \tag{1}$$

provided one of the nodes is in community r and the other node is in community r'. It is important to note that constraints on either the θ_i values or the matrix P are required in order to make the model identifiable. Here we use the constraint

$$\sum_{i=1}^{|V|} \theta_i \times I\{node\ i\ \in community\ r\} = |V_r| \quad (2)$$

where $i = 1,2, \dots, |V_r|$ and $r = 1, 2, \dots, R$. Each element in the adjacency matrix C_t at time t is generated as a Poisson random variable with mean $\lambda_{i,j} = \theta_i \theta_j P_{r,r'}$, i.e.,

$$C_t(i,j) \sim Poisson\left(\lambda_{i,j}\right). \quad (3)$$

3.2 Priebe's Scan Method

Priebe's scan method (Priebe et al., 2005) is based on three sequences of statistics, the degree, $O_{t,i}^0$, and the first two neighborhood sizes, $O_{t,i}^1$ and $O_{t,i}^2$, across all nodes. The moving window based scan statistics are calculated using a two-step standardization procedure within moving windows of width 20. The standardized statistics of the degree and the first two neighborhood sizes, $O_{t,i}^{k*}$, are calculated in the first standardization procedure as

$$O_{t,i}^{k*} = \frac{O_{t,i}^k - avg(O_{t,i}^k)}{\max(sd(O_{t,i}^k), 1)}$$

where

$$avg(O_{t,i}^k) = \frac{1}{20}\sum_{j=1}^{20} O_{t-j,i}^k\,, and$$

$$sd(O_{t,i}^k) = \left[\frac{1}{19}\sum_{j=1}^{20}\{O_{t-j,i}^k - avg(O_{t,i}^k)\}^2\right]^{\frac{1}{2}}$$

for $k = 0, 1, 2$, and $i = 1, 2, \ldots, |V|$ using the data from the previous 20 time periods. Then the statistics M_t^0, M_t^1, and M_t^2 are calculated as the respective maxima of $O_{t,i}^{0\,*}$, $O_{t,i}^{1\,*}$, and $O_{t,i}^{2\,*}$, i.e., $M_t^k = \max\{O_{t,1}^{k\,*}, O_{t,2}^{k\,*}, \ldots, O_{t,|V|}^{k\,*}\}$ and $M_t^{0\,*}$, $M_t^{1\,*}$, and $M_t^{2\,*}$ are then calculated from M_t^0, M_t^1, and M_t^2 values using the following equations in a second round of standardization:

$$M_t^{k\,*} = \frac{M_t^k - avg(M_t^k)}{\max(sd(M_t^k), 1)},$$

where

$$avg(M_t^k) = \frac{1}{20}\sum_{j=1}^{20} M_{t-j}^k \text{ , and}$$

$$sd(M_t^k) = \left[\frac{1}{19}\sum_{j=1}^{20}\{M_{t-j}^k - avg(M_t^k)\}^2\right]^{\frac{1}{2}},$$

for $k = 0, 1, 2$. The network is monitored over time using statistics $M_t^{0\,*}$, $M_t^{1\,*}$, and $M_t^{2\,*}$. The presence of an anomaly is signaled when one or more of the three standardized maxima statistics exceeds the signaling threshold. Zhao et al. (2017) have shown that 4, rather than the value 5 recommended by Priebe et al. (2005), is a reasonable signaling threshold. Thus, a signal is generated at time t if $M_t^{0\,*} \geq 4$, or $M_t^{1\,*} \geq 4$, or $M_t^{2\,*} \geq 4$. Although Priebe's scan method was proposed for binary data, we also use it with count data where we show it can be more effective.

4 Simulations

We assumed that the social network has a known community structure in our simulations, the unscaled values of parameters θ_i, denoted as θ_i', were first generated from a Pareto distribution,

$$\theta_i' \sim iid\ PAR(m = 1, s = 3)$$

for $i = 1, \ldots, |V|$, where m denotes the scale parameter and s denotes the shape parameter. The Pareto distribution was chosen for its ability to represent heterogeneity among the nodes, mainly that the node degrees are distributed according to a power law, as observed in many empirical networks. These values were then scaled to θ_i using the following the constraint,

$$\theta_i = \frac{\theta_i'}{\sum_{i=1}^{|V_r|} \theta_i'} \times |V_r|$$

Since nodes within the same community have a higher propensity to communicate with one another than to nodes from a different community, the values of the diagonal elements of P, $P_{r,r}$ were set larger than the values of the off-diagonal elements, $P_{r,r'}$. The ratio between $P_{r,r}$ and $P_{r,r'}$ was kept constant at 2:1 in the simulations. Moreover, low $\lambda_{i,j}$ values result in networks with low communication levels, or sparse networks. Non-sparse networks with high communication levels would be generated by use of high $\lambda_{i,j}$ values.

We performed our simulations using data aggregated on five different levels, $1\times$, $2\times$, $5\times$, $10\times$, and $20\times$. In order to establish the two moving windows, Priebe's scan method requires a minimum of 40 adjacency matrices. Thus,

to accommodate the $20\times$ aggregation level and assure that there are sufficient data for Priebe's scan method, a total of $T = 860$ adjacency matrices were generated sequentially for each round of simulation. The communication data contained in C_t are aggregated over the time between time period t and $t-1$. We refer to this level of aggregation as the $1\times$ aggregation level, i.e. no additional aggregation. When network data are aggregated, communications are combined over a longer period of time. In this case, the sum of all communications over several time periods would be recorded and reported at once. For any aggregated observation, communication data contained in the aggregated adjacency matrix will equal to $C_t + \cdots + C_{t-w+1}$, where w is the aggregation level. To simulate aggregation levels higher than $1\times$, we sum a sequence of adjacency matrices to represent the network communications at that level. For instance, at $5\times$ aggregation, the matrices C_t to C_{t-4} are summed to form a new adjacency matrix, such that

$$C_t^{(5)}(i,j) = \sum_{f=t-4}^{t} C_f\,(i,j), \qquad t = 5, 10, 15 \ldots$$

Both network count data and binary data are included in our investigation. Binary data are obtained by converting the communication counts in C_t into binary form. The corresponding binary adjacency matrix is denoted by B_t. In our simulation, the conversion threshold was chosen to be 1. However, it is important to keep in mind that the conversion threshold can be selected to be larger. As a result, elements in B_t that are equal to 1 indicate the presence of communication between the corresponding pair of individuals, and elements that are equal to 0 indicate no communication between the corresponding pairs within the given aggregation period. Network binary data aggregated at a higher level are obtained by summing

consecutive blocks of B_t matrices. As such any elements equal or greater than 1 in the sum are converted to 1. Including network binary data in our investigation permits us to study how aggregation at different levels affect the performance of Priebe's scan method when limited information is available. By converting count adjacency matrices C_t into binary adjacency matrices B_t, we guarantee that both network count and binary data are generated from the same simulations and the results from the use of the two data types are comparable.

We will briefly introduce the monitoring phases associated with network surveillance methods before explaining our strategy for simulating network anomalies. In network surveillance, the time period where the network communication baseline is established is referred to as *Phase I*. The *Phase I* in our simulations contain a minimum of 40 adjacency matrices due to the two moving windows required by Priebe's scan method. *Phase II* refers to the period where real time monitoring occurs. During *Phase II,* a signal is given if communication behavior in the network appear to have deviated significantly from the baseline level. If the level of communications in the network continues to be close to the baseline level, then the network is said to be *in-control*. In contrast, networks in which the communication level changes significantly are determined to be *out-of-control*. The source of the significant change in communication is typically a network anomaly. During this phase, network anomalies are generally characterized as an event for which the communication among a subset of nodes has increased or decreased by a significant amount. More often than not, anomalies are associated with an increase in communication rather than a decline, and often center around particular communities. When network data are aggregated, anomalies can occur at the beginning, middle, or end of an aggregation period. However, it is important to note that the earliest any network surveillance method can signal in *Phase II* is at the end of an aggregation period.

Our strategy is to introduce anomalies based on a single community during *Phase II*. We refer to $P^0_{r,r}$ and $P'_{r,r}$ as the baseline and anomalous propensity of communication within community r, respectively. For simplification, community #1 is chosen as the location for the anomaly in our simulations. The nodes within community #1 follow baseline behavior during *Phase I* and the initial period of *Phase II*, where $P_{1,1} = P^0_{1,1}$. When an anomaly is introduced, the propensity of communication among the nodes within community #1 increases significantly, and $P_{1,1} = P'_{1,1}$. From Equation (1) we see that an increase in $P_{1,1}$ corresponds to an increase in the expected number of communications between nodes i and j, corresponding to an increase in $C_t(i,j)$, where $i,j \in |V_1|$. We refer to the time period at which the anomaly is introduced as the *shift time*. In our simulation study, the increase in $P_{1,1}$ is sustained, meaning once $P_{1,1}$ shifts to $P'_{1,1}$ it does not return to $P_{1,1}$. This persistent increase in $P_{1,1}$ represent a permanent surge in the propensity of the communication within community #1 in the network. Greater surges are simulated by larger shifts in $P_{1,1}$.

Our simulations include aggregation levels of $1\times$, $2\times$, $5\times$, $10\times$, and $20\times$, as well as shifts of $P'_{1,1} = (1+s) \times P^0_{1,1}$, where s denotes the relative shift magnitude and $s = 0.5$ or 1.5. We did not change the degree heterogeneity parameters θ_i in the simulation. A total of 860 adjacency matrices are generated, each containing the network data aggregated at $1\times$ level at time period $t, t = 1, 2, \ldots, 860$. Correspondingly, the total number of adjacency matrices at each aggregation level are $430, 172, 86,$ and 43 for $2\times$, $5\times$, $10\times$, and $20\times$ respectively. Starting at the shift time, $P_{1,1}$ increases to $P'_{1,1}$ and given that nodes i and j are both in community #1, elements $C_t(i,j)$ in C_t increase considerably. The anomaly is introduced at shift times ranging from $t = 840$ to $t = 859$ to simulate situations where anomalies can occur at any point within an aggregation period.

Taking $20\times$ aggregation as an example, $C_{t=42}^{(20)}$ is calculated from 1 baseline and 19 anomalous C_t's if the anomaly occurs at $t = 841$. Conversely, 5 baseline and 15 anomalous C_t's make up $C_{t=42}^{(20)}$ when the anomaly occurs at $t = 845$. For the aggregation levels higher than $1\times$, an equal number of simulations are performed for all possible shift times associated with an aggregation period. In our simulations, adjacency matrices generated at $t = 1 \ldots 839$ make up *Phase I*, while *Phase II* consists of those generated at $t = 840 \ldots 859$.

We recognize that in addition to the detection performance, detecting network anomalies in a timely manner is equally important. We limit our simulation to 20 networks post-shift to account for the negative impact that high aggregation levels can have on signaling delays. This value was chosen because the highest aggregation level in our simulation is at $20\times$. If Priebe's scan method correctly detected the presence of an anomaly within $t = 20$ time periods after the shift at the $1\times$ aggregation level, it is designated as a success, otherwise as a failure. For instance, at the $5\times$ aggregation level, a successful detection is defined by Priebe's scan method signaling the presence of the network anomaly within four aggregated periods. The proportion of the total simulations which resulted in the successful detection of the anomaly is referred to as the detection rate. By using the detection rate as a performance metric, we quantify the amount of information lost from using binary data instead of using count data. We assume that network anomalies are persistent with sustained shifts in $P_{1,1}$. Because any anomalous communication increases are summed due to aggregation and as a result amplifying the signal, detection rates are generally expected to be higher for data that are aggregated at higher levels if given ample time. In addition to the detection rate, Priebe's scan method generates a false alarm if it signals for an anomaly within $t = 20$ time periods after the shift time given that no shift is introduced in the simulation. We

refer the proportion of simulations which resulted in least one false alarm as the false alarm rate, or FAR.

5 False Alarm Rates

One can always lower the signaling threshold to increase the sensitivity of Priebe's scan method. In order to be sure that the cause of any performance differences can only be attributed to different aggregation levels, we need to be sure that FAR does not change dramatically under different levels of aggregation.

Using the signal threshold of 4, we generated $T = 860$ networks of order $|V| = 20$ with nodes evenly distributed into two communities. The P matrix has diagonal values ranging from $P^0_{r,r} = 0.2$ to $P^0_{r,r} = 5$, representing various sparsity levels. The off-diagonal values equal to $P^0_{r,r'} = 0.5 \times P^0_{r,r}$. Shift times ranging from $t = 840$ to $t = 859$ were designated, at which point there was no shift introduced and $P_{1,1}$ is kept at the baseline value $P_{1,1} = P^0_{1,1}$ for the entire simulation. A total of 1000 simulations were performed at each aggregation level. The FAR results from the use of network count data and binary data are shown in Figures 1 and 2, respectively. We see that the estimated FARs are nearly zero for all aggregation levels and sparsity levels. These results show that any variations in anomaly detection performance should only be caused by aggregation levels. The FAR is not affected by the use of different network data types or the network sparsity levels. The FAR results that we obtain from networks with different sparsity levels and network data types are comparable.

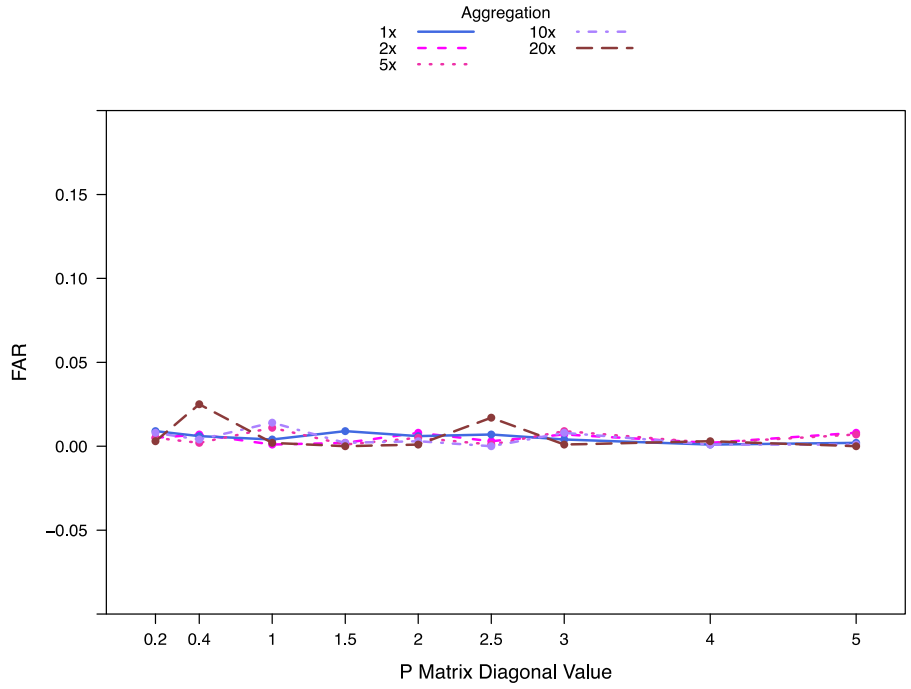

Figure 1. FAR against $P_{r,r}^{0}$ of $|V| = 20$ with two communities for aggregation levels at $1\times, 2\times, 5\times, 10\times$, and $20\times$. Use of count data.

Figure 2 shows that there is virtually no difference in FAR between the use of count or binary data. It is important to note that practitioners can adjust the signaling threshold to achieve the preferred false alarm rate. Establishing the false alarm rate or some other measure of baseline performance is crucial in designing a monitoring method.

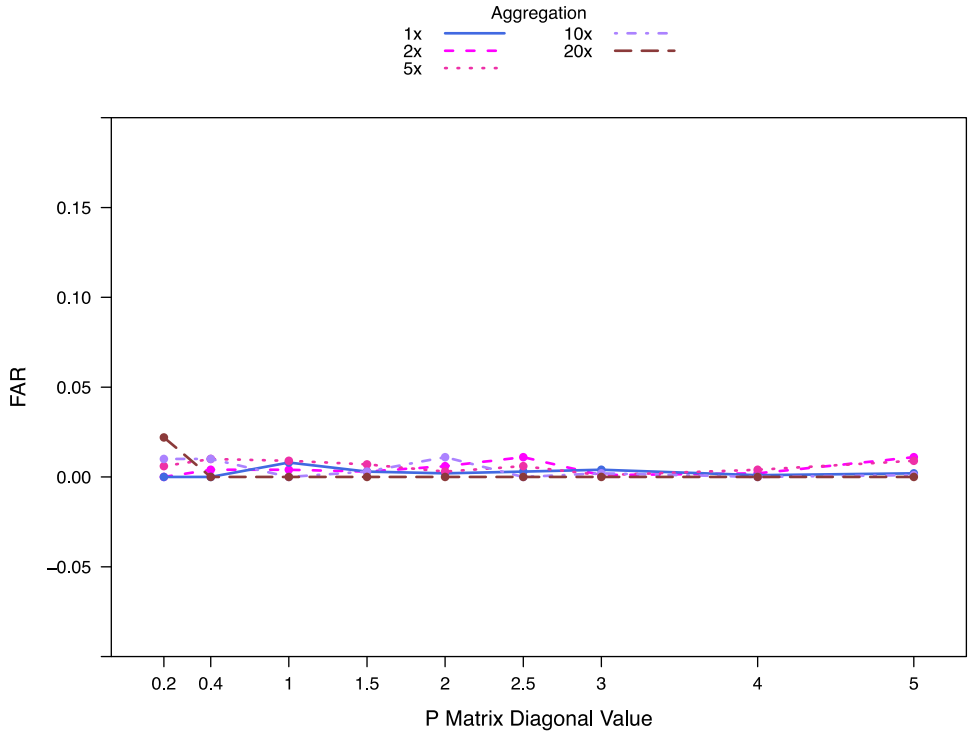

Figure 2. FAR against $P^0_{r,r}$ for $|V| = 20$ with two communities for aggregation levels at $1\times, 2\times, 5\times, 10\times$, and $20\times$. Use of binary data.

6 Evaluation

We evaluated the effect of aggregation using both count and binary data, combined with a range of network sparsity levels and shift magnitudes. We primarily employ networks of orders $|V| = 20$ and $|V| = 50$ due to computational constraints.

6.1 Network sparsity level

The effect of aggregation can vary significantly between a relatively non-sparse network and a sparse network. To explore the relationship between aggregation and sparsity levels, we performed simulations using a wide range of $P^0_{r,r}$ and $P^0_{r,r'}$ values. Anomalous shifts to be detected were fixed at $P'_{1,1} = (1 + s) \times P^0_{1,1}$, where $s =$

0.5 and 1.5. The networks simulated were of order $|V| = 20$ with two communities, each containing 10 nodes. The signaling threshold was set to 4.

Figure 3 shows the detection rates at different network aggregation levels for count data with the shift $P'_{1,1} = (1 + 0.5) \times P^0_{1,1}$. Each point represents the average detection rate from 2000 simulations. Figure 3 shows that Priebe's scan method performs poorly for the case when $P^0_{r,r} = 0.2$, where detection rates are less than 10% regardless of aggregation level. For all aggregation levels, detection rates increase as $P^0_{r,r}$ increases. Overall, the performance of Priebe's scan method is better for relatively non-sparse networks than sparse networks. When evaluating detection rates for data aggregated at different levels, the results show that detection rates increase from 0% to nearly 30% in case of $1 \times$ aggregation and from 5% to 90% for data aggregated at the $10 \times$ level. This positive trend is consistent at all aggregation levels. The most substantial improvement in performance is seen when the aggregation level increases from $2 \times$ to $5 \times$, with an increase of 50 percentage points in detection rates in some cases. Increasing the aggregation level from $5 \times$ to $10 \times$ also results in modest improvement in performance where the detection rate can increase by 20 percentage points. The performance of Priebe's scan method is generally better with respect to detection rate for data aggregated at higher levels, with the exception of $20 \times$. When $P^0_{r,r} \geq 1.0$ and specifically for data aggregated at $20 \times$, the detection rates are 20 percentage points lower than that of data aggregated at the $10 \times$ level. When $P^0_{r,r} \geq 1.5$, detection rates for data aggregated at the highest level are 10 to 20 percentage points lower than those for data aggregated at $5 \times$. At the $20 \times$ aggregation, the performance is generally lower than that of data aggregated at the $10 \times$ level. In cases where the network is relatively non-sparse where $P^0_{r,r} \geq 1.5$, the performance can be lower than of the $5 \times$ aggregation level.

For levels of aggregation higher than $20\times$, the performance will continue to deteriorate.

Individual adjacency matrices are essentially summed into one when aggregation is employed, as a result the severity of an anomaly is amplified by significantly increasing the magnitude of communication surge. Increasing the aggregation level leads to higher degree of amplification and further benefit the performance of Priebe's scan method. However, in most situations the anomaly needs to be detected quickly after its occurrence so that practitioners can investigate the anomaly as soon as possible. It is unrealistic to expect unlimited data and time to be available after any anomaly occurs in the network. The ability to detect anomalies in a timely manner is a quality that is as important as the detection rate performance for any network surveillance method. By design, our simulations extend past $t = 840$ by 20 adjacency matrices aggregated at the $1\times$ level to imitate these practical scenarios. Thus, as the aggregation level increases the number of available network adjacency matrices diminishes quickly. At the $20\times$ aggregation level, Priebe's scan method only has a single opportunity to detect the presence of an anomaly within the specified 20 $1\times$ time periods whereas the method has ten, four, and two chances for detection when data are aggregated at the $2\times$, $5\times$, or $10\times$ levels, respectively. Therefore, aggregating data at the highest $20\times$ level has a distinct disadvantage. Additionally, the need to wait until the end of an aggregation period to observe a signal from an anomaly can lower performance as well. We will explore conditional signal delays in detail in Section 7. It is worth noting that the performance drop is not a direct result of aggregation at the $20\times$ level, but a consequence of the limited data availability post-shift and the lack of signaling opportunity. We expect to see similar performance decrease for the $10\times$ or $5\times$

aggregation levels if our simulations extended to 10 or 5 adjacency matrices post-shift.

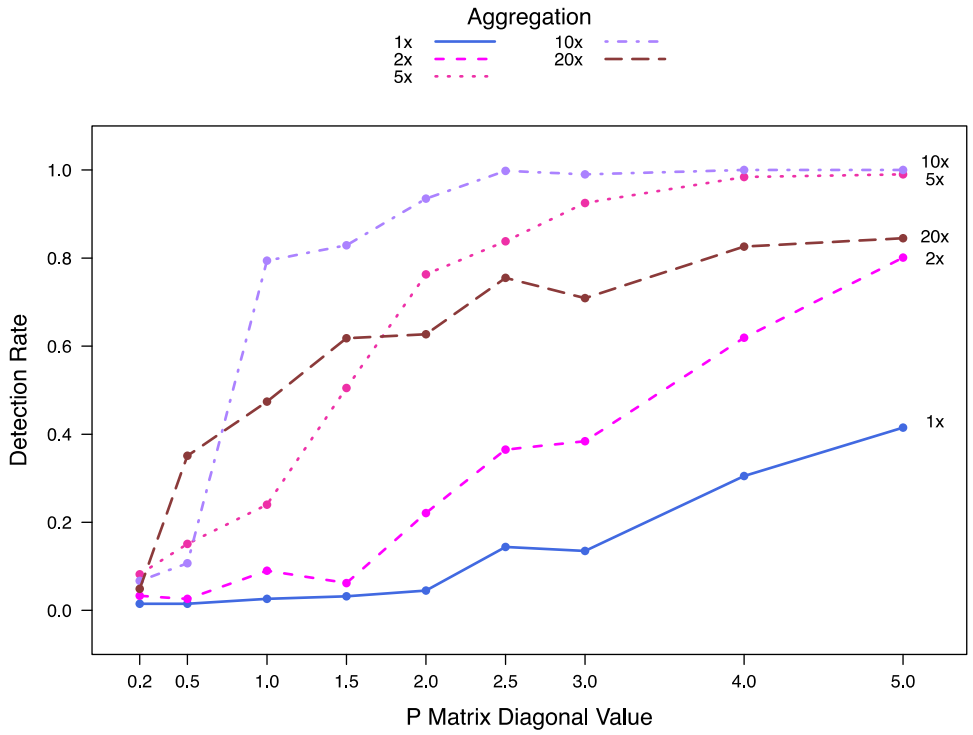

Figure 3. Detection rates against $P^0_{r,r}$ for communication data of $|V| = 20$ with two communities for aggregation levels at $1\times, 2\times, 5\times, 10\times$, and $20\times$. $P'_{1,1} = (1 + 0.5) \times P^0_{1,1}$. Use of count data.

We also studied the detection rates when binary data are used instead of count data. Although not shown here, the performance of Priebe's scan method is significantly lower compared to the results shown in Figure 3. Priebe's scan method was unable to detect any anomalies regardless of the sparsity or aggregation level. Detection rates were nearly 0% for all simulations, the poor performance can be attributed by the $P^0_{r,r}$ values used in the simulations. Our particular simulations using $P^0_{r,r} = 0.2$ or larger result in relatively non-sparse B_t matrices for the networks generated, in other words, at any given time t we often have $B_t(i,j) = 1$. When an

anomaly occurs, a high percentage of the elements in the B_t matrices representing communications among the anomalous nodes remain unchanged. Priebe's scan method performs poorly this case since there are no discernable differences between the baseline and anomalous B_t matrices. This is a significant penalty for using binary data if count data are available.

We performed identical simulations with significantly more severe anomalies using $s = 1.5$. Figure 4 shows a similar pattern compared to Figure 3 where the detection rates increase as network sparsity level decreases for the shift where $P'_{1,1} = (1 + 1.5) \times P^0_{1,1}$. The performance is significantly better for all aggregation and sparsity levels in these simulations due to the larger shift. When comparing aggregation levels, the largest improvement is again seen when aggregation level increases from $2\times$ to $5\times$. In this case, detection rates increase by as much as 30 percentage points for relatively sparse networks. For data aggregated at $20\times$, the detection rates, similar to Figure 3, are lower than those for $10\times$ by 10 percentage points. It is, however, higher than $5\times$ across all sparsity levels. We see that increasing the level of aggregation can improve performance up to a certain point, but the diminishing data availability from too much aggregation quickly leads to a decrease in performance.

We obtained detection rates in the context of binary data as well for the shift of $P'_{1,1} = (1 + 1.5) \times P^0_{1,1}$. The overall detection rates improved slightly due to the larger shift, ranging between 2% to 10% across all aggregation and network sparsity levels. However, for both shift magnitudes, the results showed that there is a significant loss of information in situations where binary data is used in place of count data.

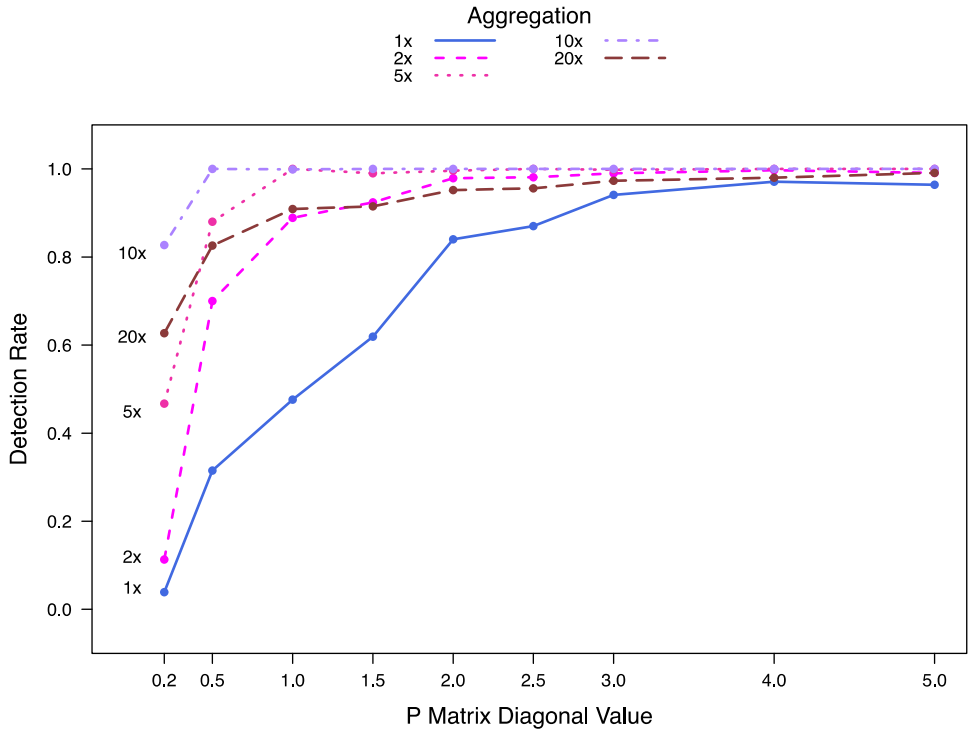

Figure 4. Detection rates against $P^0_{r,r}$ for communication data of $|V| = 20$ with 2 communities for aggregation levels at $1\times, 2\times, 5\times, 10\times$, and $20\times$. $P'_{1,1} = (1 + 1.5) \times P^0_{1,1}$. Use of count data.

In addition to simulating networks of order $|V| = 20$, we applied the same investigation by simulating larger networks of order $|V| = 50$ and $|V| = 100$. Figure 5 illustrates the detection rates at different network aggregation levels for count data with the shift $P'_{1,1} = (1 + 0.5) \times P^0_{1,1}$ where the networks are of order $|V| = 50$. The networks contain two equal sized communities each made up of 25 nodes. Each point represents the average detection rate from 2000 simulations. Figure 5 shows a similar detection rate pattern compared to that of the $|V| = 20$ networks. Priebe's scan method again performs poorly for $P^0_{r,r} = 0.2$, at which point the detection rates for the $1\times, 2\times$ and $5\times$ aggregation levels are less than 10% and are at 50% at the $10\times$ and $20\times$ aggregation levels. The performance improves

across all aggregation levels as network sparsity level decreases. When comparing the aggregation levels, the most significant improvement in performance is now observed when the aggregation increases from $1\times$ to $2\times$. When aggregating at the $2\times$ level, the detection rates are as much as 40% higher than aggregating at $1\times$. The detection rates for data aggregated at $20\times$ are consistently lower than that of data aggregated at the $10\times$ level for all sparsity levels. When $P_{r,r}^{0} \geq 1.0$, detection rates for data aggregated at the $10\times$ level are lower than those for data aggregated at $5\times$.

Figure 5 shows marginally different trends for detection rates at the aggregation levels studied. A node within a larger $|V| = 50$ network can be in communication with a considerably higher number of other nodes compared to a node within $|V| = 20$ network at any given time t. There is a significant increase in baseline communications among nodes during *Phase I* since an individual node has 49 potential targets with which to establish a connection. For shifts of identical magnitude, it would be more difficult for Priebe's scan method to detect the presence of an anomaly in *Phase II*. For the $|V| = 50$ networks in our simulations, aggregating at the $5\times$ level presents the optimal balance between amplifying the severity of the anomaly and giving the method ample opportunity for detection. Although not shown here, our simulations also revealed that for the shift $P'_{1,1} = (1 + 1.5) \times P^{0}_{1,1}$, the detection rates for $|V| = 50$ networks are comparable to those in Figure 4 with similar differences. The detection rates from the use of count data in networks with order of $|V| = 100$ are shown in Figure 6, the shift is set at $P'_{1,1} = (1 + 0.5) \times P^{0}_{1,1}$. The networks contain four equal sized communities with 25 nodes each. We observe similar trends in detection rates when compared to Figures 4 and 5. However, regardless of the aggregation level or network sparsity, the detection rates never reach 100%. Clearly, the relative order of a community within a network

affects the performance of the method. This is not dissimilar to the findings in Zhao et al. (2017), where the performance of Priebe's scan method is poor for small anomalous sub-networks.

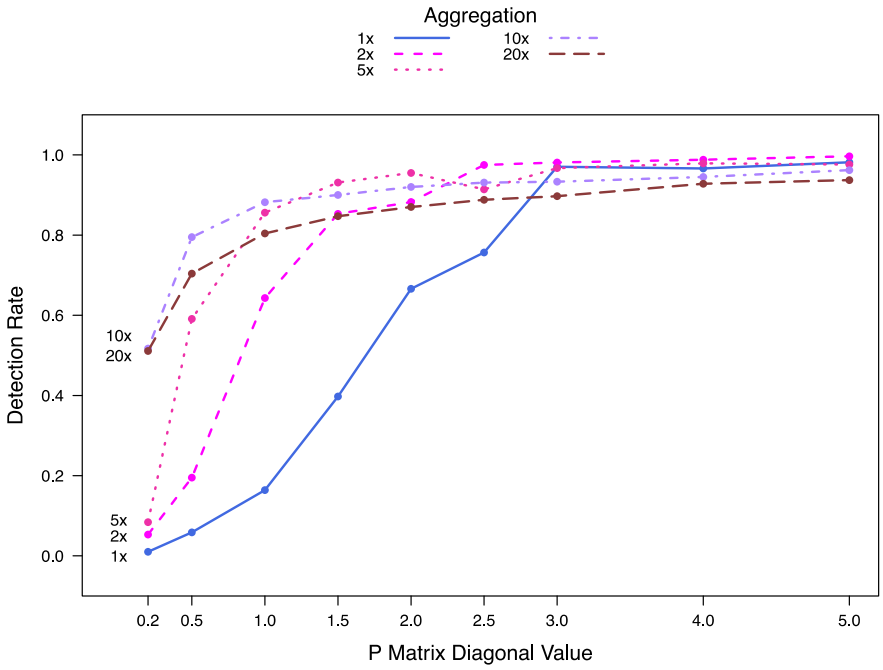

Figure 5. Detection rates against $P^0_{r,r}$ for communication data of $|V| = 50$ with 2 communities for aggregation levels at $1\times, 2\times, 5\times, 10\times$, and $20\times$. $P'_{1,1} = (1 + 0.5) \times P^0_{1,1}$. Use of count data.

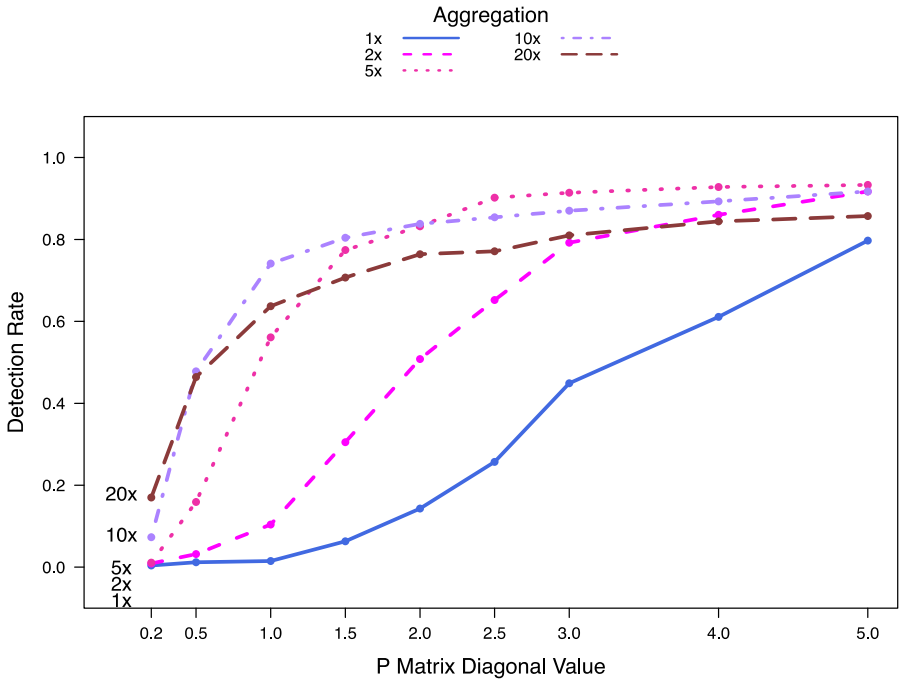

Figure 6. Detection rates against $P^0_{r,r}$ for communication data of $|V| = 100$ with 4 communities for aggregation levels at $1\times, 2\times, 5\times, 10\times$, and $20\times$. $P'_{1,1} = (1 + 0.5) \times P^0_{1,1}$. Use of count data.

To summarize, our results show that when aggregation level is at $20\times$, there is a notable drop in performance. The decrease in performance can be attributed to the lack of availability of post-shift adjacency matrices and the long aggregation period. While aggregating network data can improve detection performance, for practitioners, the optimal aggregation level should be chosen by considering factors such as data availability and fast detection. For comparisons of methods based on count and binary data, the performance of Priebe's scan method is much higher when data contain the actual number of pairwise communications in the network. Count data offer practitioners a clear advantage over binary data. We will examine the performance from the use of network binary data in more detail in Section 6.3.

6.2 Shift magnitudes

The severity of the anomaly is simulated by the shift magnitude, which is another factor that can significantly alter the performance of Priebe's scan method. We study how shift magnitudes change the effects of aggregation in this section. Using the shift $P'_{1,1} = (1+s) \times P^0_{1,1}$, where $s = 0.05, 0.25, 0.5, 1,$ and 1.5, networks of order $|V| = 20$ were generated, again with the nodes evenly divided into two communities. The baseline diagonal value of P is fixed at $P^0_{r,r} = 1$.

Figure 7 shows detection rates for network count data aggregated at various levels at different values of s, where s represents the relative shift magnitude in the value of $P^0_{1,1}$ in our simulations. When $s \leq 0.25$, the detection rates are low for all aggregation levels. When $s > 0.25$, aggregation improves the performance. Generally, the detection rates are higher for larger shifts. Studying the results from using data aggregated at different levels reveals that increasing the aggregation level from the $2\times$ to the $5\times$ level leads to the biggest improvement in detection rates, echoing the results in Figures 3 and 4. Similar to results discussed Section 6.1, when $0.50 \leq s \leq 0.75$, detection rates for the $20\times$ aggregation level are 20 to 40 percentage points lower than those of the $10\times$ level. When $s \geq 1$, the detection rates for the $20\times$ aggregation level are 10 to 20 percentage points lower than those for data aggregated at $5\times$. Aggregating at the $20\times$ level shows lower performance resulting from the limited post-shift data availability and the single chance for detection. While results from the use of binary data are not shown here, detection rates were nearly 0% regardless of the aggregation level after converting the networks to binary form, reinforcing that using binary data can lead to a significant loss of information.

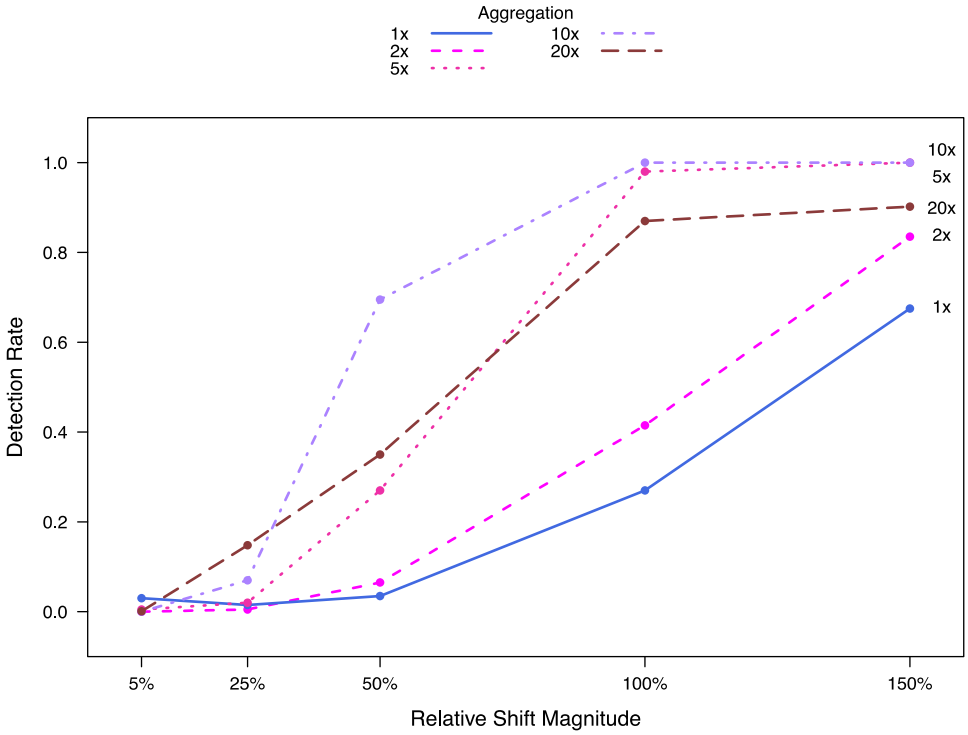

Figure 7. Detection rates against shift magnitudes for network data for aggregation levels at $1\times, 2\times, 5x, 10\times$, and $20\times$. $P_{r,r} = 1$. Use of count data.

6.3 Revising Simulations for Network Binary Data

The results shown in Section 6.1 and 6.2 suggest that converting network count data into binary data result in considerably lower detection rates. Binary data present a distinct disadvantage compared to count data in network surveillance applications. In addition, the results also demonstrated that the simulation settings used are not appropriate for studying the effect of aggregation using network binary data. Our simulations did not provide the information needed to gain any insights on the effect of aggregation. Because the conversion from network count data to binary data is not linear, the same anomaly severity specified in the simulation for count data does not translate to binary data form in a one-to-one manner. In this section, we compare the expected change in B_t pre- and post-shift in order to find applicable simulation settings for binary data. Specifically, network densities are used for comparison,

where it is defined as the proportion of the potential connections in a network that are actual connections.

We generated 2000 networks of order $|V| = 20$ each with two equal sized communities, where $P'_{1,1} = (1 + 1.5) \times P^0_{1,1}$, and $P'_{1,1} = (1 + 4) \times P^0_{1,1}$. Figure 8 shows the distribution of network densities for B_t pre- and post-shift and at two different sparsity levels. In both cases, the network densities are not affected by the sparsity level. The pre- and post-shift values only differ by 15%, not the +150% as specified in the simulation for the $P_{1,1}$ component. For Figure 9, which shows the distribution of network densities for B_t pre- and post-shift and at two different sparsity levels in the case of $P'_{1,1} = (1 + 4) \times P^0_{1,1}$, the difference in the densities are only slightly more than 20%. In order to detect anomalies reliably with binary data, the difference between the pre- and post-shift densities needs to be significantly higher. As such, $P^0_{r,r}$ has to be much smaller while s needs to be much larger. In other words, the severity of the anomalies need to be much stronger with the use of network binary data in order for Priebe's scan method to detect the anomalous signals.

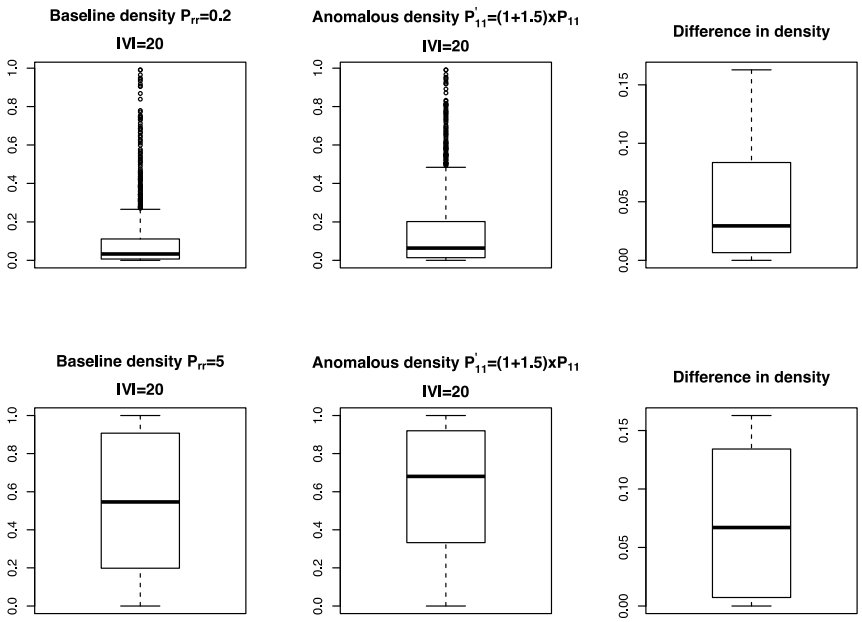

Figure 8. Network densities pre- and post-shift. Binary networks of $|V| = 20$ with 2 communities. Shift magnitude is +150%.

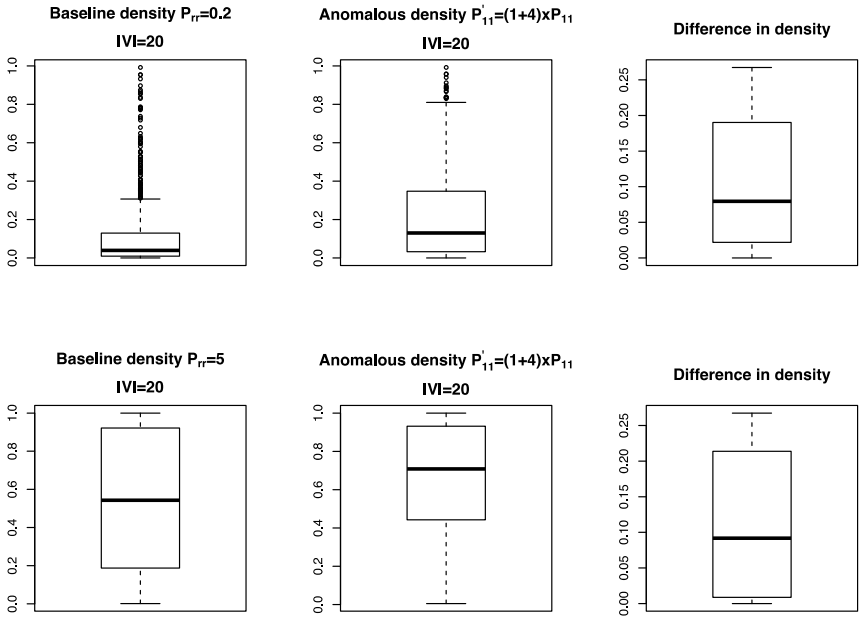

Figure 9. Network densities pre- and post-shift. Binary networks of $|V| = 20$ with 2 communities. Shift magnitude is +400%.

Consequently, we simulated networks with $P_{r,r}^0$ ranging from 0.02 to 0.2. Anomalies are simulated with much larger shift magnitudes for the propensity of communication within community #1, with $P'_{1,1} = (1 + s) \times P_{1,1}^0$. We set a fixed value for $P'_{1,1} = 1$ and this results in s ranging from 5 to 50 with the various values of $P_{1,1}^0$. A total of 2000 simulations were performed for each combination of s value and aggregation level. By adopting a wide range of s values in our simulations, we are able to vary the difference between the pre- and post-shift network density significantly. Figure 10 shows the detection rates by aggregation levels at different $P_{r,r}^0$ values. When $P_{r,r}^0 = 0.02$, Priebe's scan method is able to detect the presence of an anomaly with a reliable consistency. Detection rates range from 80% to 100% depending on the aggregation level. When $0.04 \leq P_{r,r}^0 \leq 0.12$, detection rates decrease significantly by as much as 50 percentage points. The same trend is observed regardless of the aggregation level. This is because as $P_{r,r}^0$ increases, there are more baseline node-to-node interactions, i.e. $B_t(i,j) = 1$ pre-shift. The change in the network densities post-shift in turn is small, making anomaly detection more difficult. When comparing different aggregation levels, we see that there is little difference in detection rates among different aggregation levels. Aggregation at any level x results in $B_t^{(x)}(i,j) \geq 1$, which is then converted to $B_t^{(x)}(i,j) = 1$ using our conversion threshold. The resulting $B_t^{(x)}$ matrices are very similar regardless of the aggregation level. This similarity negates the effect of aggregation in which anomalous communications are summed to amplify anomaly severity. We performed identical simulations using network count data as well. While not shown, the performance with the use of count data for all s values is much higher with detection rates that are generally at 95% and frequently above 99%.

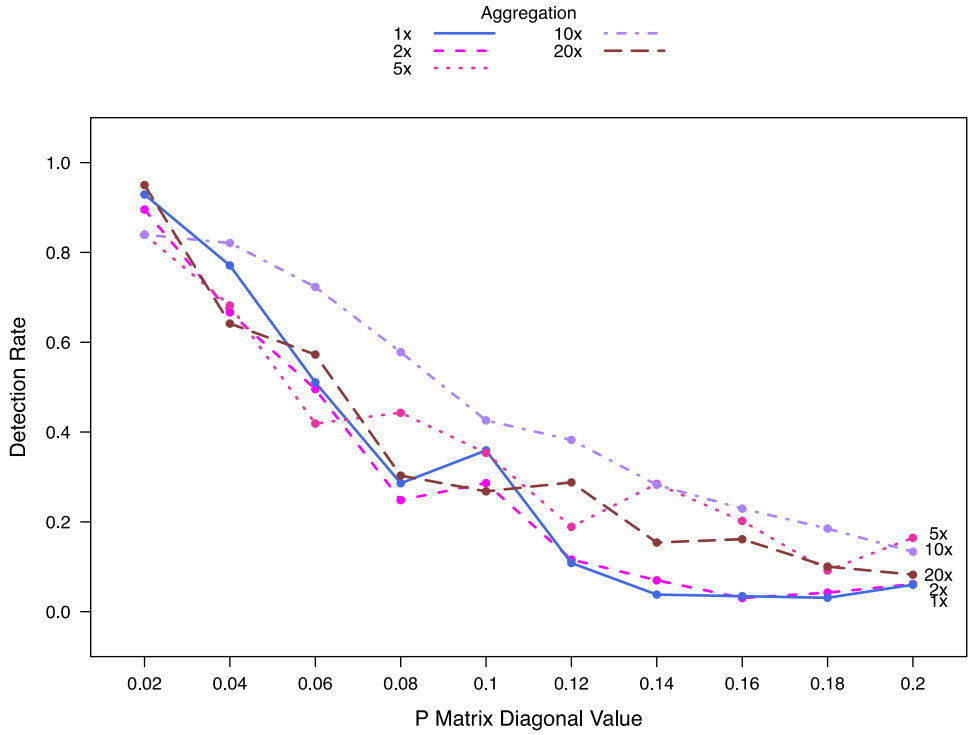

Figure 10. Detection rates against $P_{r,r}$ for $|V| = 20$ networks with two communities for aggregation levels at $1\times, 2\times, 5\times, 10\times$, and $20\times$. Use of binary data.

7 Distribution of Conditional Signal Delay

Given an anomaly has been successfully detected, we refer the time it takes for the method to signal the *conditional signal delay*. While the main goal for any network surveillance method is to detect the presence of a network anomaly, additionally the *conditional signaling delay* caused by high levels of aggregation is also a serious concern. Practitioners are often interested in both detection performance and the time it takes for the method to signal. We address the inherent delay due to aggregation by studying the conditional delays when monitoring aggregated count data in this section. Performing the same simulations as shown in Section 6, we generated networks of order $|V| = 20$ with nodes evenly distributed into two communities. While simulating anomalies using $P'_{1,1} = (1 + 1.5) \times P^0_{1,1}$, we calculated the *conditional signal delay* from the proportion of 1000 simulations where the

anomalies have been detected successfully. Figures 11 shows the distribution of *conditional signal delay* for 1 × aggregation level. We see that there is generally no delay in signaling. Detection is almost immediate because there is no need to wait until the end of a long aggregation period. A large majority of signals occur within two time periods following the shift.

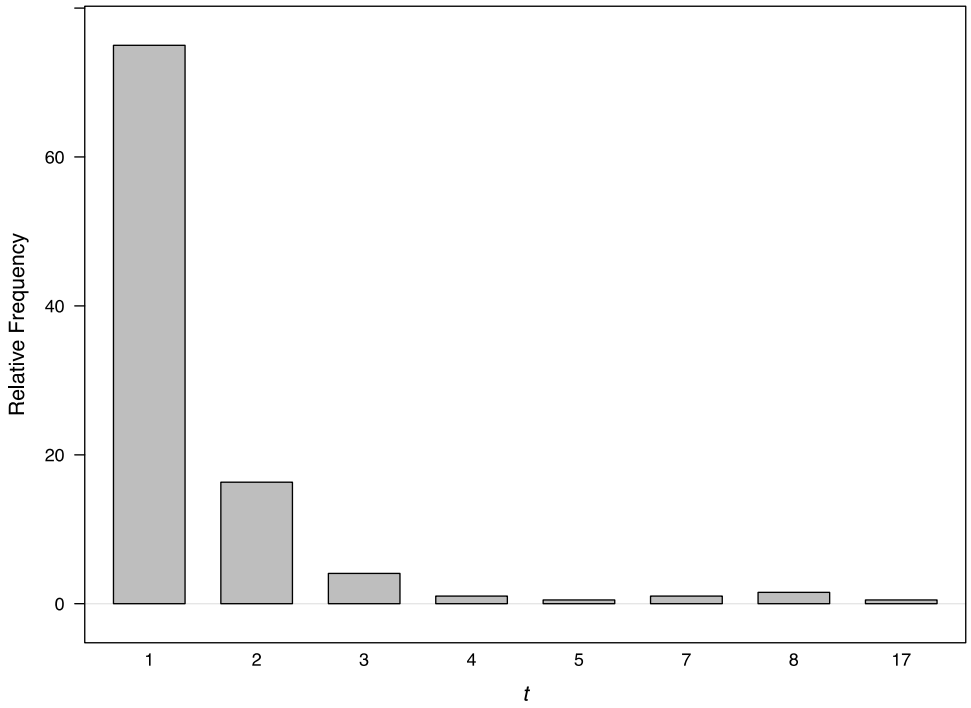

Figure 11. Distribution of conditional signal delay for network counts data aggregated at 1 × level. $P'_{1,1} = (1 + 1.5) \times P^0_{1,1}$.

While we did not observe much delay in anomaly detection using data aggregated at the 1 × level when signals occur, greater signaling delay is unavoidable at higher aggregation levels. Figure 12 shows the *conditional signal delay* for 2 ×, 5 ×, 10 ×, and 20 × aggregation levels. At the 2 × aggregation level, anomaly detection took slightly longer in a few cases. For even higher aggregation levels, we see that signaling progressively took longer as aggregation levels increased. At the 20 × aggregation level, over half of the signals occur more than

ten time periods following the shift. Though not included in our investigation, aggregation levels that are higher than $20\times$ will cause even greater signaling delays when signals occur.

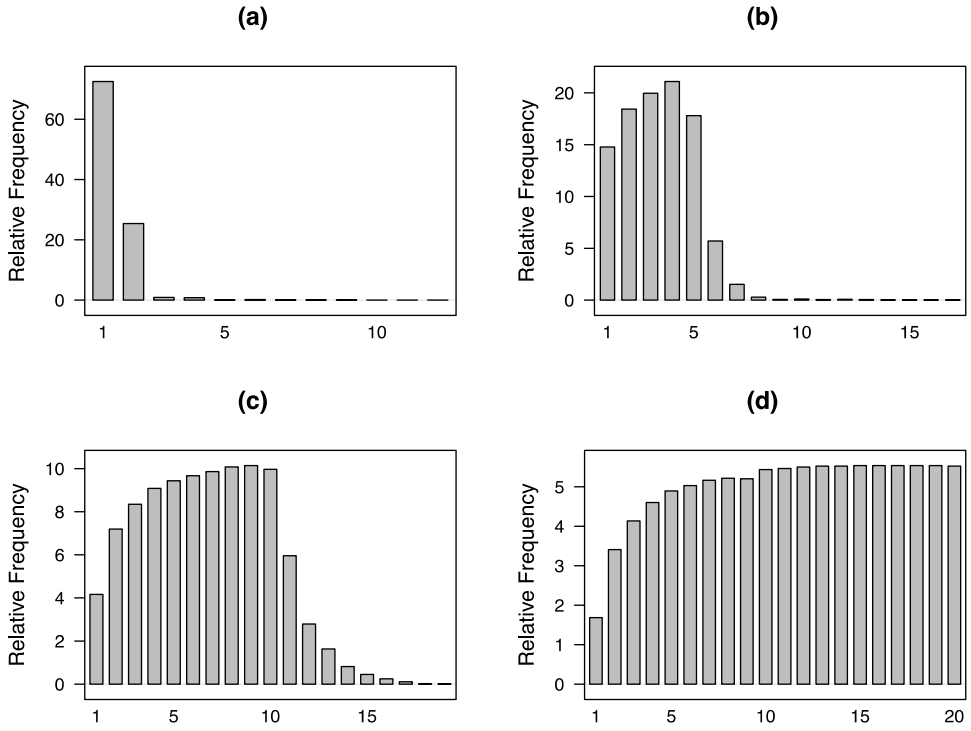

Figure 12. Distribution of conditional signal delay for network counts data aggregated at (a) $2\times$ (b) $5\times$ (c) $10\times$ and (d) $20\times$, level. $P'_{1,1} = (1 + 1.5) \times P^0_{1,1}$.

It is important to stress that for reliable detection with Priebe's scan method, a large amount of the aggregated data must be anomalous. Depending on when an anomaly occurs within an aggregation period, performance could be negatively affected, especially for the $20\times$ aggregation level specifically studied in our simulations. With $20\times$ aggregation, there are 20 possible shift times within an aggregation period. If the shift occurs during the latter part of the aggregation period, only a small portion of the aggregated data will be anomalous. Figure 13 shows how shift times affect the performance with respect to each aggregation level. Both $5\times$ and $10\times$ aggregation levels show lower detection rates if anomalies occur during

the second half of the aggregation period compared to the beginning of the aggregation period. Moreover, we see a slight increase in detection rates immediately after the performance drop. The recovery in performance is attributed to the fact that the aggregation period immediately after is made-up entirely of anomalous data, giving the method another chance to signal. At 20 × level, we see that if anomalies occur at or after the 12th adjacency matrix in a block of 20 within the aggregation period, the performance starts to decrease, and continues to do so until the detection rate drops to 25%. In this case, Priebe's scan method only has one chance to signal within 20 time periods due to the high level of aggregation, and cannot detect the anomalies reliably when less than 40% of the data within the aggregation period are anomalous. The performance will further decrease for even higher aggregation levels.

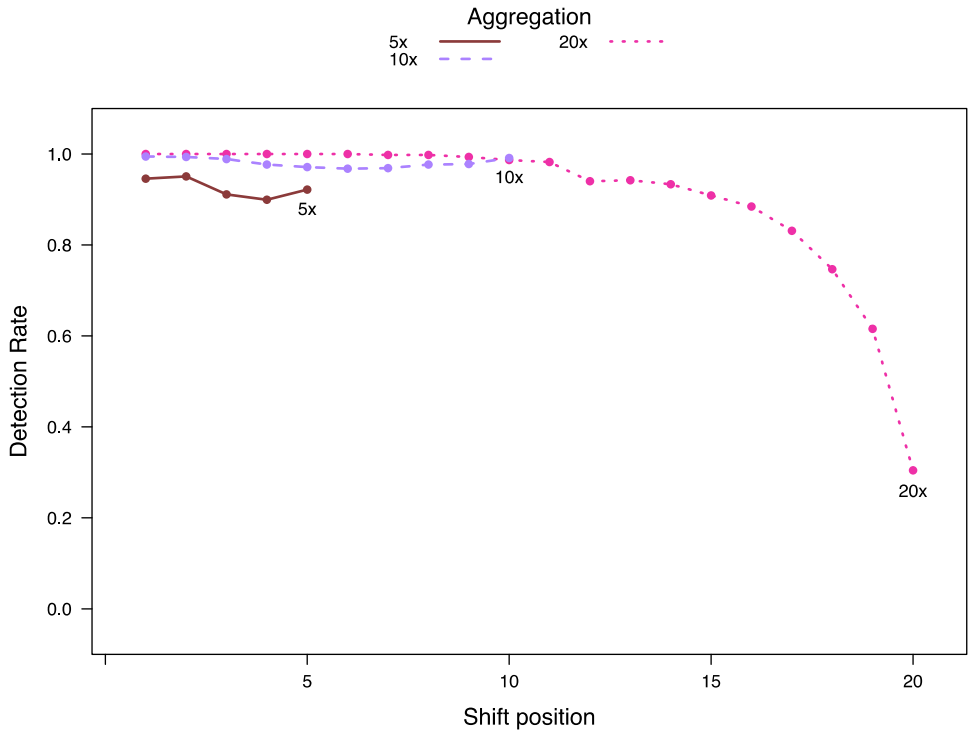

Figure 13. Detection rates against shift time for communication data of $|V| = 20$ networks with 2 communities for aggregation levels at $1 \times, 2 \times, 5 \times, 10 \times$, and $20 \times$. $P'_{1,1} = (1 + 1.5) \times P^0_{1,1}$. Shift only occurs in $P_{1,1}$. Use of count data.

8 Conclusions

We have demonstrated how aggregating network data at different levels affects the performance of Priebe's scan method. Our results show that aggregating network data can be beneficial to the performance of Priebe's scan method up to a certain point and then becomes detrimental. Additionally, the simulation studies show that there is a significant loss of information in situations where network binary data are used in place of count data.

When communication counts are recorded, aggregation results in sums of individual adjacency matrices and amplifies the severity of the anomaly. We saw in our specific examples that the most significant increase in detection rates was observed when increasing the aggregation level from $2\times$ to $5\times$ level. Increasing the aggregation level from $5\times$ to $10\times$ also show modest increases in detection performance in our examples. Our simulation outcomes reveal that there is a noticeable drop in performance at the $20\times$ aggregation level. Using our specific simulation settings coupled with this aggregation level, the performance is substantially lower than that of the $10\times$ aggregation level, and when the networks are relatively sparse it is even lower than that with the $5\times$ aggregation level.

By studying the conditional signal delay and the relationship between detection performance and shift time, we are able to attribute the poor performance to two causes. First, when there is a limited time window for detection post anomaly, as the aggregation level increases the number of available adjacency matrices diminishes quickly. At a certain level, it is likely that the network surveillance method used only has a single opportunity to detect the presence of an anomaly within a specified time window. Given the specific simulation conditions in our investigation, Priebe's scan method often failed to detect the anomaly using data aggregated at the $20\times$ level within 20 time periods after the shift. Second, the precise time at which an anomaly occurs is unknown to practitioners. Depending on

the aggregation level, there are a number of possible shift times within an aggregation period. If an anomaly occurs during the latter part of an aggregation period, only a small portion of the aggregated data are anomalous. Successful immediate detection is made very difficult in these situations. Practitioners must wait until the end of the next aggregation period to determine if an anomaly is present. The results showed that if an anomaly occurs at one of the last five possible shift times within an aggregation period at the $20\times$ level, detection rates can change by as much as 50 percentage points. While our simulations show lower performance at the $20\times$ aggregation level, we believe that the main cause of the performance degradation is the signaling delay resulting from aggregating network data.

Furthermore, when comparing network binary data against count data, the performance of Priebe's scan method is much lower when binary data are used to represent whether pairwise communications in the network have surpassed a certain level. In contrast, use of count data offers practitioners a clear advantage over binary data in that much more information on network communications is retained.

By comparing the expected network densities pre- and post-shifts, we were able to adjust the settings of our simulations to dramatically increase the severity of the anomalies. The results from the resulting simulations showed declines in performance as the shift magnitude decreased with the use of binary data. However, unlike network count data, there is little difference in detection rates among different aggregation levels. Aggregated binary data can be very similar regardless of the level used. This similarity negates any benefit of aggregation.

We expect the overarching theme of our results to generalize to other situations. We recommend the use of network count data over binary data. Count data retains the actual number of node-to-node communications rather than only showing if pairwise communications have surpassed a certain level. Additionally, aggregating count data involves combining the communication counts cumulated

over the aggregation period. Information on the difference in communication counts among pairs of individuals over time are retained, where this information can be lost by the use of binary data. Moreover, we recommend the use of aggregation at the appropriate level for network surveillance applications. The specific aggregation level used depends upon the signal delay that can be tolerated in the monitoring situation.

It would be impractical to study all of the possible cases of data aggregation in network surveillance. We consider one model and one method in our study, but other models and methods could be studied. Moreover, we assume that the network anomalies are based on a single community, however, the nature or type of the anomalies can be defined differently and studied as well. We want to reiterate that our investigation is based on simulations that represent just a fraction of the possible scenarios in network surveillance where aggregated data are used. For example, the network anomalies are assumed to be persistent in our study. For cases in which the anomalies are transient in nature, higher levels of aggregation are likely to yield worse performance. In addition, the deterioration in detection performance from the use of network binary data warrants further investigation.

References

Benneyan, J. C. (1998). Statistical Quality Control Methods in Infection Control and Hospital Epidemiology, Part II: Chart Use, Statistical Properties, and Research Issues. Infection Control and Hospital Epidemiology, 265-283.

Dubrawski, A., and Zhang, X. (2010). The Role of Data Aggregation in Public Health and Food Safety Surveillance. Biosurveillance: Methods and Case Studies, 161.

Erdős, P., and Rényi, A. (1959). On Random Graphs. I. Publicationes Mathematicae, 6: 290– 297.

Fricker, R. D. Jr. (2013). Introduction to Statistical Methods for Biosurveillance: With an Emphasis on Syndromic Surveillance. Cambridge University Press.

Karrer, B., and Newman, M. E. J., (2011). Stochastic Blockmodels and Community Structure in Networks. Physical Review E, 83(1): 016107.

Kulldorff, M. (2001). Prospective Time Periodic Geographical Disease Surveillance Using a Scan Statistic. Journal of the Royal Statistical Society: Series A (Statistics in Society), 164(1), 61-72.

Priebe, C. E., Conroy, J. M., Marchette, D. J., and Park, Y. (2005). Scan Statistics on Enron Graphs. Computational & Mathematical Organization Theory, 11(3): 229-247.

Savage, D., Zhang, X., Yu, X., Chou, P., and Wang, Q. (2014). Anomaly Detection in Online Social Networks. Social Networks, 39: 62-70.

Schuh, A., Woodall, W. H., and Camelio, J. A. (2013). The Effect of Aggregating Data When Monitoring a Poisson Process. Journal of Quality Technology, 45(3), 260-272.

Schuh, A., Camelio, J. A., and Woodall, W. H. (2014). Control Charts for Accident Frequency: a Motivation for Real-Time Occupational Safety Monitoring. International Journal of Injury Control and Safety Promotion, 21(2), 154-162.

Wartenberg, D. (2001). Investigating Disease Clusters: Why, When and How?. Journal of the Royal Statistical Society: Series A (Statistics in Society), 164(1), 13-22.

Wilson, J. D., Stevens, N. T., and Woodall, W. H. (2016). Modeling and Estimating Change in Temporal Networks Via a Dynamic Degree Corrected Stochastic Block Model. arXiv preprint arXiv:1605.04049.

Woodall, W. H., Marshall, J. B., Joner Jr, M. D., Fraker, S. E., and Abdel-Salam, A. S. G. (2008). On The Use and Evaluation of Prospective Scan Methods

for Health-Related Surveillance. Journal of the Royal Statistical Society: Series A (Statistics in Society), 171(1), 223-237.

Woodall, W. H., Zhao, M. J., Paynabar, K., Sparks, R., and Wilson, J. D. (2017). An Overview and Perspective on Social Network Monitoring. IISE Transactions, 49(3), 354-365.

Zhao, M. J., Driscoll, A. R., Fricker, R. D., Jr., Sengupta, S., Spitzner, D. J., and Woodall, W. H., (2017) Performance Evaluation of Social Network Anomaly Detection Using a Moving Window Based Scan Method. Submitted for publication.

Chapter 5

Conclusion and Future Work

We presented our research in two parts in this dissertation. In Chapter 2, we evaluated the performance of the moving window based scan method proposed by Priebe et al. (2005). In general, using existing social network data and illustrating performance with a case study for a proposed method is not sufficient. We recommend using simulated social networks as a general tool with which to evaluate the performance of social network monitoring methods. We have also investigated the effects of aggregation levels in social network surveillance in Chapter 4. Depending on the availability of post-anomaly data and the need for fast detection, aggregation at certain levels can improve anomaly detection performance when network count data are available. Additionally, we recommend network count data to be used, as it retains considerably more information on network communications than binary data.

Our simulations in Chapter 2 showed that Priebe's scan method performs poorly when the anomalous sub-network size was a small percentage of the size of the entire network, which is often the type of situation in which a user is most interested. Even when the sub-network was as large as 10% of the entire network the method had difficulty detecting anomalies. In scenarios where the SNR value was low, the anomalous signal generated did not represent significant deviations from normal network behavior. Our results have shown that Priebe's scan method is not suitable for detecting anomalous events with low SNR values. To improve performance, we proposed several modifications to Priebe's scan method, including lowering the signaling threshold, implementing a variable moving window, and lagging the moving window. Additional results demonstrated that these improvements can significantly improve the method's performance.

We continued our investigation into social network surveillance in Chapter 4, for which network data aggregation was the main focus. By employing the DCSBM in conjunction with Priebe's scan method, we were able to simulation network anomalies based on intrinsic community structures that are present in social networks. Using both network count and binary data aggregated at various levels, we were able to demonstrate how aggregating network data at different levels affects the performance of Priebe's scan method. Our results showed that aggregating network data can be beneficial to the performance of Priebe's scan method up to a certain point and then becomes detrimental. The simulation studies showed that there is a significant loss of information in situations where network binary data is used in place of count data.

Our research considers a single monitoring method. While we studied network surveillance using both the Erdős–Rényi and the DCSBM models, other models and methods could be used as well. We want to reiterate that our investigation is based on simulations that represent a fraction of the possible scenarios in network surveillance. For example, the network anomalies are assumed to be persistent in our study. For cases in which a network anomaly occurs within a large network that contains thousands of nodes and only involves a few nodes; or that an anomaly that involves nodes across multiple communities. In addition, social networks are often complex entities that represent realistic interactions among individuals, and could be better characterized by other models. All of these factors can serve as the focus for future studies.