Distribution Agreement

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

Signature:

_____ _____

Isabel Chen Date

Centrality measures and contagion on temporal networks

By

Isabel Chen
Doctor of Philosophy

Mathematics

Michele Benzi, Ph.D.
Advisor

James Nagy, Ph.D.
Committee Member

Vicki Stover Hertzberg, Ph.D.
Committee Member

Accepted:

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

Date

Centrality measures and contagion on temporal networks

By

Isabel Chen

Advisor: Michele Benzi, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Mathematics
2016

Abstract

Centrality measures and contagion on temporal networks

By Isabel Chen

The objective of this dissertation is to study the relationship between network-based centrality measures and epidemic outcome. Determining the key players in contagion processes can inform disease-prevention strategies. We analyze a time-stamped, person-to-person contact network based on human mobility movements within a busy, urban hospital. Movement patterns identified a small number of locations as hubs of activity. Linear algebraic techniques were used to compute a recently proposed temporal centrality measure applied to the empirical network; comparisons with traditional centrality measures were performed to determine if the inclusion of temporal information provides additional insights. Linear regression techniques were employed to describe the relationships between the quantities of interest. We find that while temporal centrality can at times identify key players not captured by traditional measures, it does not necessarily outperform non-temporal measures with respect to predicting epidemic outcome. Strategic removal of connections between highly central nodes resulted in an exponential decrease in the structural connectivity of the network, but this did not translate to a reduction in epidemic outcome. We conclude that contagion on temporal networks is extremely robust to changes in the network, and while network-based centrality can help to identify key players in an epidemic process, more work needs to be done to build an epidemic-containment strategy based on the information afforded by network-based analyses.

Centrality measures and contagion on temporal networks

By

Isabel Chen

Advisor: Michele Benzi, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Mathematics
2016

Acknowledgments

This work could not have been completed without the support of my advisor, Prof **Michele Benzi**. Thank you for guiding me through the completion of this dissertation.

To Prof **Vicki Hertzberg**, in addition to your role in my committee, many thanks for providing the interesting dataset and kickstarting this project.

To Prof **James Nagy**, thank you for being on my dissertation committee, and for your kindness over the years.

I owe much of my training in statistical analysis to Prof **Howard Chang**, who kindly shared his expertise, as well as good humor.

Thanks also to Prof **Vǒjtech Rödl** for inspiration and encouragement, especially during my early years at Emory University.

To the other staff and faculty of Emory's Mathematics & Computer Science Department and Laney Graduate School, thank you for your support in making my experience at Emory as pleasant as it was.

Finally, much gratitude to my friends and family, particularly my husband, **Pascal Philipp**, who patiently rode all the ups and downs with me. Thanks for sharing the ride.

Without you all, nothing would be possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The presence of infectious agents in a confined space brings substantial risks of cross infection. The aim of this thesis is to make use of network analysis to better understand contagion processes within such spaces. We base our study on the interactions of people in an Emergency Department (ED) of a hospital in the Midtown area of Atlanta, GA [50].

Disease spread within a healthcare facility is a legitimate concern. According to the CDC, on any given day, about one in 25 hospital patients has at least one healthcare-associated infection (HAI). In 2011, there were an estimated 722,000 HAIs in acute-care hospitals in the US; 75,000 patients with HAIs died during their hospitalizations. Examples of HAIs include central line-associated bloodstream infections, catheter-associated urinary tract infections, surgical site infections and gastrointestinal illness. Our study focuses on the spread of infectious diseases (such as influenza) via close-proximity person-to-person contact, which form up to 16% of reported HAIs.

According to [83], the influenza virus can be transmitted via three route categories: (i) large droplets expelled by an infectious person through coughing, sneezing or talking directly into the eyes, nose or mouth of a susceptible person; (ii) smaller, aerosolized particles inhaled by a susceptible person; and (iii) hand-to-face self-inoculation after touching a contaminated

person, surface or object. Apart from (iii), transmission of the influenza virus requires close-proximity contact between an infectious agent and a susceptible person. In the rest of this work, we will model the problem with such interactions between agents in mind.

The importance of understanding the mechanics of disease transmission cannot be understated, especially in this day and age where the majority of the world's population live in dense, urban centers. Early mathematical models of epidemics were developed in [38, 68], with major contributions to the modern theoretical framework provided by [3, 7] and [43]. An underlying assumption of these models is that interactions between agents are well-mixed, in other words, every individual has an equal chance of spreading the disease to every other. Utilizing contact network structure allows this assumption to be relaxed: instead of assuming that interactions between agents occur homogeneously and at random, contact between agents can be modeled based on their underlying network structure [25, 42, 48, 66]. Consequently, disease spread depends not only on the population size and the infection/recovery rate, but also on the (typically non-homogenous) connectivity of agents as determined by their positions in the network.

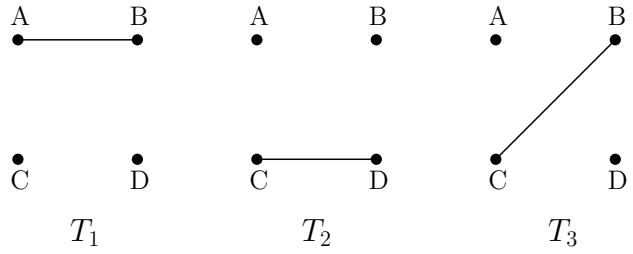
Static networks assume that contacts are permanent and unchanging over time. This simplifying assumption is sufficient in the case when the population under study is relatively stable [9, 42], and when the rate of infection is high [40, 54]. Analyses using the framework of static networks have been useful in the context of disease ecology [10], animal health problems [40], as well as public health problems relating to SARS, walking pneumonia, influenza and gonorrhea [54, 12, 11, 24, 2].

While static network approximations are sufficient in some cases, the following studies have shown that temporal features such as duration, frequency, temporal ordering and concurrency (temporal overlap) of contacts, can significantly influence the spread of disease and should be incorporated into epidemiological models if possible (see [8] for a comprehensive discussion). Regular contact patterns that are long in duration can result in a reduction

in epidemic size [63]; [72] showed that exposure-dependent transmission models give rise to epidemic outcomes that are significantly different from models in which all contacts (regardless of length) are equally likely to transmit the disease; inclusion of contact timing and duration can have varying effects in different settings [77]; in [73] it was shown that increased frequencies of repeated contacts reduces epidemic size, and this effect is stronger when the number of contacts is small, or transmission probability is low. It is therefore evident that temporal information can provide insights that may otherwise be overlooked.

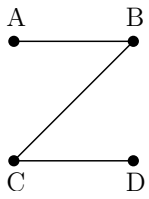
We further illustrate the importance of the temporal order of contact sequences: a person can only pass on the disease to others after becoming infectious. For illustration, consider the following time-evolving network on four nodes shown in Figure 1.1, with node A being the only infectious node at time T_1. An edge between two nodes is representative of some form of disease-transmitting contact between them.

Figure 1.1: An example of a dynamic network



For simplicity, we assume that an infected person is immediately infectious. There is a temporal path $A \to B \to C$ which means that the disease can potentially spread from A to both B and C. On the other hand, it is impossible for D to become infected, because the contact between C and D occurs *before* C has contact with B. Note that the static version of the network (Fig 1.2) loses such information: without the temporal dimension, all nodes, including D, are reachable from A. Work in [30] and [81] have shown that the non-constancy of contact structure can have a significant impact on disease spread, and should therefore be incorporated into the disease model if such information is available.

Figure 1.2: The aggregated version of Figure 1.1



The study of empirical contact data has revealed patterns that differ from a priori contact assumptions [56, 63], which highlights the need to better understand real-world data. Recent technological advances have made high-dimensional data available for study, and thorough analyses have been performed on empirical temporal networks based on person-to-person interactions at conferences [19, 74], sexual contacts [65], human mobility patterns based on cell phone data [33], and diary-based data on social contacts at home and at the workplace [63]. Since patterns of human behavior depend greatly on the context, results based on one particular network may not generalize easily to other networks. We hope to contribute to this growing body of work, by performing an in-depth study of person-to-person interactions within the ED of a busy, urban hospital.

In Chapter 2, the construction of a contact network based on the observed data is described. Here, the nodes represent people and the edges represent their interactions. The temporal network is represented as a sequence of adjacency matrices on the same set of nodes (as in Figure 1.1), where each matrix represents the connections present at the corresponding time-step. The time-evolution of the network is thus captured by the appearance and disappearance of edges over time. We point out that there are no edges linking nodes across time-steps, and therefore the network under study is neither a multilayer [46] nor a multiplex [20] network.

An important objective is to identify key players in the infection process, and *centrality measures* are a natural choice to consider for this task. Centrality measures are numerical scores associated with each node in the network, which are computed based solely on the

underlying network structure. These numerical scores can then be used to *rank* the nodes, quantifying a notion of how 'central', or how 'important', a node is relative to the other nodes in the network. Many different centrality measures have been used and proposed [18, 26, 59], each one capturing a particular notion of 'importance'. It is often the case that a node may rank highly with respect to one measure, but not with another measure. Such differences are in and of themselves interesting, and may elucidate properties of the node in question that are not immediately obvious. It is therefore instructive to use multiple centrality measures as probes to shed light on the connectivity structure of the network under study.

In this work we focus on temporal generalizations of *walk-based* centrality measures. Examples of walk-based measures are node degree, Katz centrality, communicability and eigenvector centrality. These are in contrast to *path-based* centrality measures, such as betweenness and closeness centrality, which are typically more computationally expensive. Walk-based centrality measures can be expressed concisely as matrix functions on the matrix-representation of the network, leading to novel applications of matrix theory to network science.

The relationship between centrality and epidemic outcome has been studied in [6, 14, 45, 69] and [71], but the networks under study did not include temporal information. It has also been argued that while some centrality measures are able to identify highly influential nodes, they do not accurately quantify [14, 49], and may even underestimate [79], the spreading power of the vast majority of nodes which are not highly influential. Indeed, work in [72] suggests that the importance of highly connected individuals is 'strongly overestimated' when the duration of contacts is not taken into account. Alternative spreading power metrics such as the accessibility [80, 78] and expected force [49], which extend centrality by incorporating spreading dynamics, have been shown to have stronger correlation with epidemic outcome than traditional centrality [14, 47, 71]. Before rejecting centrality

measures completely as a means to understand contagion on networks, we seek to answer the question: will incorporating *temporal information* into centrality metrics improve their explanatory and/or predictive power in relation to epidemic outcome?

We are motivated by examples in [52] and [75] which show that centrality measures based on time-aggregated versions of temporal networks (as shown in Fig. 1.2) fail to adequately capture important nodes. The work in [65] suggests that temporal correlations in network data should not be underestimated and that consequently, detecting important individuals based on temporal structure may have a significant impact on targeted intervention strategies. In [69] it was shown that while centrality measures on the time-aggregated network (such as node degree, betweenness centrality, eigenvector centrality) can improve immunization strategies, node strength, which measures the total time exposed to others, was the most effective. These results illustrate the importance of utilizing temporal information in the context of identifying key players in the contagion process.

Our work aims to bring recently developed temporal centrality measures into the analysis. Temporal centrality measures provide a means of identifying important nodes based solely on time-dependent network structure. While temporal centrality is a relatively young field compared to its static counterpart, new measures are constantly being developed [39, 44, 75, 76]. We focus our attention on *dynamic communicability* introduced by Grindrod, Parsons, Higham and Estrada [37], which is a centrality measure that ranks nodes based on *temporal walks* on the network. This will be presented and discussed in Chapter 3. The results of dynamic communicability applied to the empirical temporal network are presented in Chapter 4. In Chapter 5 we take a closer look at the observed interactions between the nodes identified as 'central' by this method. We emphasize that the computation of centrality scores depend solely on the network structure and is completely independent of any contagion process on the network.

To relate network-based properties to contagion, we simulate an epidemic process on the

network, where the duration of contacts is used explicitly in the infection procedure. In other words, an epidemiological model is built based on *when* and *where* disease-spreading contacts occur. This is in principle similar to the mechanistic model of infection proposed in [72], and is a form of agent-based modeling described in [8]. As discussed in [8], while agent-based simulations are an alternative to analytical models of disease spread on dynamic networks (see for example [13, 64, 81]), such methods face computational and algorithmic problems, which we encountered first-hand. Nonetheless, agent-based simulations have the advantage of being easy to understand and interpret, and with the lack of a general mathematical framework that can handle a broad range of realistic dynamic networks, such an approach provides a first step in understanding how contagion processes take place on the network. In Chapter 6 we describe our approach in greater detail and present the epidemic outcomes observed on the network.

In Chapter 7, we aim to address the following two questions: First, what role does temporal centrality play in explaining epidemic outcome? Second, can temporal centrality be used to predict epidemic outcome? It has often been asserted in the literature [6, 45, 47, 49, 71] that a strong correlation between epidemic outcome (Y) and some measure, say X, will result in measure X being a good predictor for Y. The notion of correlation measures the strength of linear relationship between X and Y, and is intrinsically a goodness-of-fit measure of the dataset under consideration. Determination of predictive power, on the other hand, must be assessed on a *different* dataset. While a strong correlation is suggestive of good predictive power, correlation will not reflect the predictive power between variables with a non-linear relationship. In our work, instead of looking at raw correlations, we use multiple linear regression to study the overall effect of centrality on epidemic outcome. We recognize that centrality measures alone cannot fully explain epidemic outcome, and the regression framework allows for the inclusion of other variables into the analysis in order to gain a better understanding of the role that dynamic network centrality plays in the epidemic

process. The regression model generates coefficients that quantify the relationship between predictor and response, and it is these coefficients that are used in the prediction process, not the correlation itself. Predictions are performed on different samples of networks drawn from the same study. This provides a more accurate assessment of the predictive power of centrality in the context of epidemic spread.

Employing network centrality to explain node characteristics/behavior has been done in the social sciences [67, 70], where path-based measures, such as betweenness and closeness centrality, are typically used. Our goal in this work is two-fold: first, to examine the effectiveness of *temporal network centrality* in explaining and/or predicting epidemic outcome, and second, to highlight temporal, *walk-based* measures such as dynamic communicability, which, due to their formulation in terms of matrix functions, can be computed with greater computational ease than path-based measures. The development of other walk-based measures (apart from degree and eigenvector centrality) appears to stay within the confines of the mathematics community, and we hope that our work can help to highlight other walk-based measures to the larger network science and epidemiology community.

An alternative approach to probe the relationship between network-based centrality and epidemic outcome is presented in Chapter 8. Instead of quantifying the direct relationship between centrality and virulence, we ask the question: how is epidemic outcome affected by the removal of connections between highly central nodes? We are motivated by work in [1, 4, 5], which showed that strategically targeting highly central nodes can have a strong effect on the overall connectivity of the network. We show that the same is observed on this empirical network, and additionally seek to study the changes in epidemic outcome associated with reduction in the structural connectivity of the network.

The overarching aim of this work is to study the relationship between network-based properties (such as centrality) and epidemic outcome. While network representations grossly simplify reality, much insight can still be gleaned from such analysis. A better understanding

of the properties of key players in the context of epidemic spread can help inform more effective epidemic-containment strategies. We hope that this work provides a comprehensive approach towards a small step in achieving this aim.

1.1 Terminology

Throughout this work, the term 'network' refers to a pair $G = (V, E)$, where V represents the set of nodes (or often called 'vertices' in traditional graph theory), and $E \subseteq V \times V$ represents the set of edges. The presence of the pair $(i, j) \in E$ means that there is an edge (or 'link') between node i and node j. The number of nodes of the network, $|V|$, is denoted by n.

The network (or 'graph') G is represented by its adjacency matrix, denoted by A. This is a non-negative $n \times n$ matrix, that is, its entries are ≥ 0, with ij-th entry representing the weight of the link between node i and node j. The matrix A is often binary, in which case the ij-th entry is indicative of the presence or absence of the edge (i, j). In this work, nodes are not linked to themselves, that is, $(i, i) \notin E$ for all $i \in V$. In other words, the diagonal entries of A are always zero. Such networks are referred to as 'simple'. We also consider only *undirected* networks – these are networks where the presence of an edge $(i, j) \in E$ means that the edge (j, i) also exists. One can also consider the edge set E of an undirected network as a subset of unordered pairs of nodes. In matrix terms, this means that the adjacency matrix A is symmetric: $A_{ij} = A_{ji}$ for all $1 \leq i, j \leq n$. We denote by A^T the *transpose* of the matrix A, which is obtained by switching the rows and columns of A. More formally, $A^T_{ij} = A_{ji}$. Therefore, a symmetric matrix is one where $A = A^T$. In this work, since adjacency matrices are often time-stamped with a time-index t, we denote A^T by A'.

Analysis of the matrix representation of the network can provide much insight into its structural properties. If there exists a non-zero vector $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = \lambda\mathbf{x}$ for some scalar λ, then λ is called an *eigenvalue* of A, with corresponding (right) *eigenvector* $\mathbf{x}$.

Eigenvalues of A are therefore the roots of the characteristic polynomial of A, $\det(A - \lambda I) = 0$, where I is the $n \times n$ identity matrix and det is the determinant function. There are therefore at most n distinct eigenvalues associated with any $n \times n$ matrix. The *spectrum* of A, denoted $\sigma(A)$, is the collection of all eigenvalues of A:

$$\sigma(A) = \{\lambda : \lambda \text{ is an eigenvalue of } A\},$$

while the *spectral radius* of A is defined as

$$\rho(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}.$$

Since A is symmetric and has real entries, there exists an orthogonal matrix Q such that $Q^T AQ$ is a diagonal matrix $D \in \mathbb{R}^n$. (An orthogonal matrix Q is one where $QQ^T = Q^T Q = I$, the identity matrix. In other words, Q is invertible with inverse $Q^{-1} = Q^T$.) The diagonal entries of D are the eigenvalues of A, and it is therefore easy to see that the columns of Q are precisely the eigenvectors of A. Since Q is invertible, all its columns must be linearly independent. This means that a real, symmetric matrix A has n linearly independent eigenvectors.

Consider a connected component of the network G. This is a subgraph $G' = (V', E')$ where $V' \subseteq V$ and the edge set $E' \subset V' \times V'$ has the property that any pair of nodes in V' can be connected by a sequence of edges in E'. The corresponding adjacency matrix A' associated with G' is termed *irreducible*, and by Perron-Frobenius Theory, there exists an eigenvalue of A' such that $\lambda = \rho(A')$, that is, the largest eigenvalue of A' is positive (> 0). This is called the Perron-Frobenius eigenvalue. Furthermore, there exists an eigenvector of the Perron-Frobenius eigenvalue which has strictly positive entries. This eigenvector, say $\mathbf{x}$, can therefore be used to *rank* the nodes in V': if $\mathbf{x}_i > \mathbf{x}_j$ then node i is ranked higher than node j. This is called *eigenvector centrality*. Further explanations can be found in [59]

and [26].

The *degree* of a node i is simply the number of neighbors of i in the network, and can also be used to rank its importance relative to other nodes in the network. More formally, the degree of node i is number edges of the form $(i, j) \in E$ for any $j \neq i$, and is given by the ith row (or column) sum of A:

$$\text{degree of node } i = \sum_j A_{ij} = \Big(A \cdot \mathbb{1} \Big)_i$$

where $\mathbb{1}$ is the vector of all ones. Node degree is therefore a centrality measure which ranks nodes according to the number of direct connections they have: the larger the number of direct links, the higher the rank. One can also view a direct link or edge as a *walk* of length one connecting two nodes.

A *dynamic, time-evolving* or *temporal* network is represented by a sequence of adjacency matrices $\{A^{[t]}\}$ on the same set of nodes. The superscript t indexes the time-step, and total number of time-steps is denoted M. In this view, the time-evolution of the network is captured by the appearance and disappearance of edges over time. In the temporal setting, the analogous notion of degree can be computed by simply adding up the adjacency matrices over time, and ranking nodes based on the degree of this aggregated matrix. We call this *aggregated degree*, denoted AD:

$$\text{AD} = \left(\sum_{t=1}^{M} A^{[t]} \right) \cdot \mathbb{1}.$$

The aggregated matrix is no longer binary, and the ij-th entry counts the number of times edge (i, j) appears over the time-span $1 \leq t \leq M$.

One can also *binarize* the aggregated matrix, by setting all non-zero entries to 1. We refer to this as the *binarized, aggregated matrix*, and the degree based on this matrix ranks nodes according to the number of distinct neighbors accrued over time.

We point out that throughout this work, highly-ranked nodes are associated with small numerical rank. That is to say, the highest-ranked (or most central) node has rank 1. In this work, various centrality measures will be used to rank nodes, and comparisons across different measures and associated rankings are performed using the following metrics: Pearson correlation (denoted *pcorr*), Kendall τ correlation (denoted *kcorr*) and intersection similarity [29] (denoted *isim*). Given two lists $\mathbf{X} = (x_1, \ldots, x_n)$ and $\mathbf{Y} = (y_1, \ldots, y_n)$, let the sample mean of $\mathbf{X}$ be $\bar{x} = \frac{1}{n}\sum_i x_i$, and let the sample standard deviation of $\mathbf{X}$ be $s_x = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$. Analogous quantities are defined for $\mathbf{Y}$. Then the Pearson correlation between $\mathbf{X}$ and $\mathbf{Y}$ is given by

$$\mathrm{pcorr}(\mathbf{X}, \mathbf{Y}) := \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right).$$

It can be shown that the $\mathrm{pcorr}(\mathbf{X}, \mathbf{Y}) \in [-1, 1]$, and furthermore, that the magnitude of pcorr is an indicator of the strength of linear relationship between $\mathbf{X}$ and $\mathbf{Y}$: a value close to 1 indicates a strong linear relationship between X and Y, while a value of $|\mathrm{pcorr}| \approx 0$ suggests that there is no linear relationship between the two lists. A negative value is indicative of an inverse relationship between $\mathbf{X}$ and $\mathbf{Y}$. Note that the value of pcorr is not the same as the *slope* of the best-fitting line between $\mathbf{X}$ and $\mathbf{Y}$: if the points of $\mathbf{X}$ and $\mathbf{Y}$ are scattered closely to a horizontal line, the Pearson correlation will be close to 1, even though the slope of the best-fit line is close to 0.

Consider now a pair of observations (x_i, y_i) and (x_j, y_j), corresponding to, say, the centrality ranks of node i and node j according to two different centrality measures. The pair of observations is said to be *concordant* if both centrality measures 'agree' with respect to the relative rankings of the two nodes: either both $x_i > x_j$ and $y_i > y_j$, or both $x_i < x_j$ and $y_i < y_j$. They are said to be *discordant* if one measure ranks node i higher than node j, but the other measure does the opposite: either $x_i > x_j$ and $y_i < y_j$, or $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant. The Kendall τ

correlation coefficient is defined as:

$$\text{kcorr}(\mathbf{X}, \mathbf{Y}) := \frac{\text{number of concordant pairs - number of discordant pairs}}{\binom{n}{2}}.$$

Again, it is easy to see that kcorr $\in [-1, 1]$, and if $\mathbf{X}$ and $\mathbf{Y}$ agree perfectly, then kcorr $= 1$, and if the two lists disagree perfectly, then kcorr $= -1$.

Finally, we will compute the intersection similarity (isim) between the computed lists. Let $\mathbf{X}^{[k]} = (x_1, \ldots, x_k)$ and $\mathbf{Y}^{[k]} = (y_1, \ldots, y_k)$ be the first k elements in $\mathbf{X}$ and $\mathbf{Y}$ respectively. The *top k intersection similarity* is given by

$$\text{isim}_k(\mathbf{X}, \mathbf{Y}) := \frac{1}{k} \sum_{i=1}^{k} \frac{\left| \mathbf{X}^{[i]} \triangle \mathbf{Y}^{[i]} \right|}{2i}$$

where $\triangle$ is the symmetric difference operator between the two sets. If the two lists are identical,

$$\mathbf{X}^{[i]} = \mathbf{Y}^{[i]} \; \forall \; i \iff \mathbf{X}^{[i]} \triangle \mathbf{Y}^{[i]} = \emptyset \; \forall \; i \iff \text{isim}_k(\mathbf{X}, \mathbf{Y}) = 0 \; \forall \; k.$$

Similarly, if the two lists are disjoint, $\text{isim}_k(\mathbf{X}, \mathbf{Y}) = 1$ for all k. Therefore a small value of $\text{isim}_k(\mathbf{X}, \mathbf{Y})$ is indicative of a strong similarity between the two (top k) rankings, while a value close to 1 suggests that the rankings are quite different.

We remark that isim_k will not be used to compare lists of centrality measures, because it is based on the intersection of the lists as sets. It is unlikely that centrality measures from different lists agree numerically, even though they may be very close. Using isim_k on lists of centrality measures will likely result in a value close to 1, since the lists are probably close to being set-wise disjoint. We therefore compute isim_k only for *lists of nodes in ranked order*, and we may also compute the set-wise intersection of these lists for comparison.

Chapter 2

Data

2.1 Construction of the temporal network

We study the interactions of people in an Emergency Department (ED) of Emory University Hospital, Midtown (EUHM), in Atlanta, Georgia. The data was collected by Vicki S. Hertzberg *et. al* in 2009 [50]. The ED was divided into 95 zones, demarcated by the blue lines shown in Figure 2.1. The zones were designed in such a way that two people in the same zone are within 1 meter of each other with very high probability. Although no physical contact is guaranteed, the close proximity of any two people within a zone can be considered a potential disease-spreading contact. Such close proximity interactions are particularly pertinent to infections that are transmitted predominantly via droplets. Participants in the study wear radio-frequency (RFID) tags which track their movements in the ED. Both patients and staff were recruited to participate. Three groups of staff were present: medical doctors (MD), registered nurses (RN) and administrative staff (other). Data was collected over 35 shifts of at most 12 hours. Of these, we analyze the data obtained from 7 shifts. Throughout this work, n denotes the number of participants per shift. Shift-specific information is shown in Table 2.1. See [50] for more details on the data

collection procedure.

Figure 2.1: Emergency Department of EUHM. There are 95 zones, demarcated by blue lines.

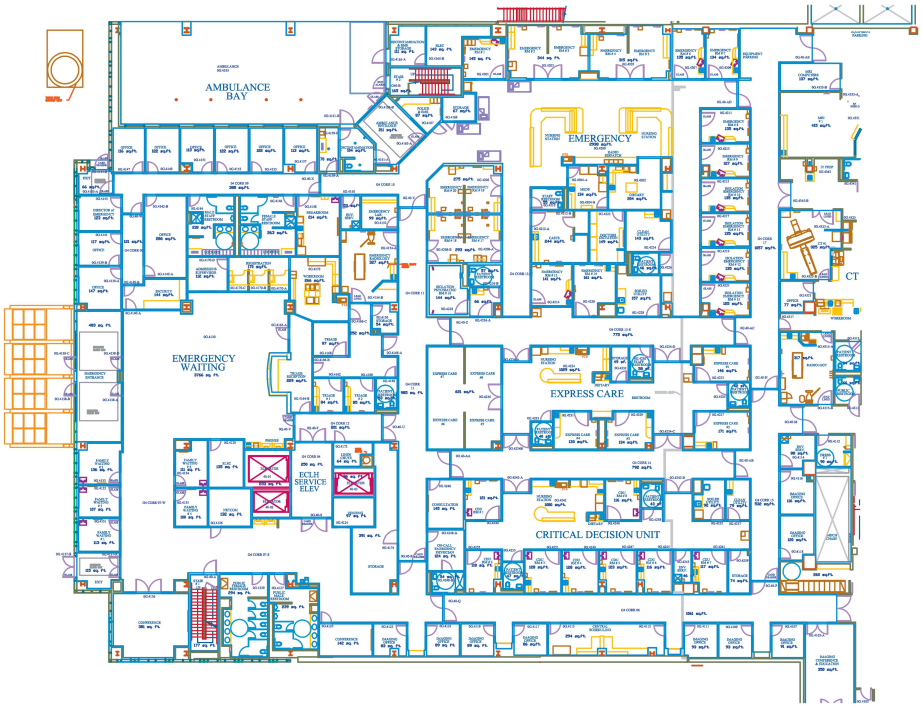


Table 2.1: Shift-specific data. The total number of participants is denoted by n. Participation rate (p. rate) is the percentage of people present in the ED who participated in the study.

Shift	n	staff	patients	total patients	p. rate (%)	shift length (h)	am/pm	weekday
1	107	33	74	98	76	11	pm	y
2	115	33	82	117	70	12	pm	y
3	89	25	64	82	78	8	am	n
4	129	34	95	108	88	12	pm	n
5	133	44	89	117	76	11.75	pm	y
6	87	26	61	77	79	8	am	n
7	126	35	91	133	68	11.67	am	n

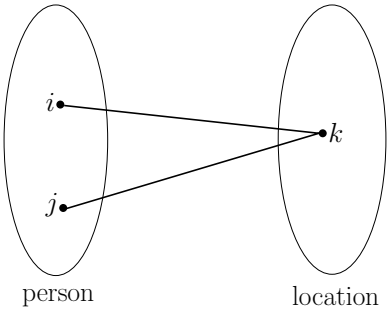
Raw data was provided was in the form of person-by-location matrices, $P^{[t]}$, one for each

second t, where

$$P_{ik}^{[t]} = \begin{cases} 1 & \text{if person } i \text{ is in location } k \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

for $t = 1, \ldots, 43202$. Since a person can be in only one place at any given second, there can be at most one non-zero entry in every row of $P^{[t]}$. (If a person is not present in the ED, the corresponding row consists entirely of zeros.) On the other hand, the column sums of $P^{[t]}$ tell us how many people were in each location at second t.

We can view the matrix $P^{[t]}$ as representing a bipartite network of the form



from which we would like to obtain the person-by-person adjacency matrix $A^{[t]}$ where

$$A_{ij}^{[t]} = \begin{cases} 1 & \text{if person } i \text{ and person } j \text{ were in the same location at time } t \\ 0 & \text{otherwise} \end{cases}$$

for $t = 1, \ldots, 43202$. To do this, for every second t we compute PP' (where for ease of notation we drop the superscript t):

$$\begin{array}{c} \\ i \\ \\ \\ j \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} & & & & k & & \end{array} \\ \begin{pmatrix} & & & & \vdots & & \\ \cdots & \cdots & \cdots & \cdots & 1 & \cdots & \\ & & & & \vdots & & \\ & & & & \vdots & & \\ \cdots & \cdots & \cdots & \cdots & 1 & \cdots & \\ & & & & \vdots & & \end{pmatrix} \end{array} \times \begin{array}{c} \begin{array}{ccccccc} & & i & & & & j \end{array} \\ k \begin{pmatrix} & & \vdots & & & & \\ & & \vdots & & & & \\ \cdots & 1 & \cdots & \cdots & 1 & \cdots & \\ & & \vdots & & & & \end{pmatrix} \end{array} = \begin{array}{c} \\ i \\ \\ \\ j \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} & & i & & & & j \end{array} \\ \begin{pmatrix} & & \vdots & & & & \\ \cdots & 1 & \cdots & \cdots & 1 & \cdots & \\ & & \vdots & & & & \\ & & \vdots & & & & \\ \cdots & 1 & \cdots & \cdots & 1 & \cdots & \\ & & \vdots & & & & \end{pmatrix} \end{array}.$$

The resulting matrix PP' is clearly square and symmetric, with (ij)-th entry equal to 1 if and only if person i and person j are in the same location (here it is at location k) during time t. Note that in this process we lose the information regarding *where* contact took place. Observe also that the resulting product is a binary $0-1$ matrix because a person cannot be in more than one place within the time-frame of one second.

By way of illustration, consider time frames of say, 30 seconds, so that a person can be in more than one location over this time period. For example, suppose person i moves through locations l, m, n within this time frame, while person j only moves through locations m and n. The product PP' now takes the form

$$\begin{array}{c} \\ i \\ \\ \\ \\ j \\ \\ \end{array}\!\!\begin{array}{c} \begin{array}{ccccccc} & & l & & & m & n & \\ \end{array} \\ \begin{pmatrix} & \vdots & & & \vdots & \vdots & \\ \cdots & 1 & \cdots & & 1 & 1 & \cdots \\ & \vdots & & & \vdots & \vdots & \\ & \vdots & & & \vdots & \vdots & \\ \cdots & \cdots & \cdots & & 1 & 1 & \cdots \\ & \vdots & & & \vdots & \vdots & \end{pmatrix} \end{array} \times \begin{array}{c} \\ l \\ \\ \\ \\ m \\ \\ n \\ \\ \end{array}\!\!\begin{array}{c} \begin{array}{ccccccc} & & i & & & & j & \\ \end{array} \\ \begin{pmatrix} & \vdots & & & & \vdots & \\ \cdots & 1 & \cdots & \cdots & \cdots & \vdots & \cdots \\ & \vdots & & & & \vdots & \\ \cdots & 1 & \cdots & \cdots & \cdots & 1 & \cdots \\ \cdots & 1 & \cdots & \cdots & \cdots & 1 & \cdots \\ & \vdots & & & & \vdots & \end{pmatrix} \end{array} = \begin{array}{c} \\ i \\ \\ \\ \\ j \\ \\ \end{array}\!\!\begin{array}{c} \begin{array}{ccccccc} & & i & & & & j & \\ \end{array} \\ \begin{pmatrix} & \vdots & & & & \vdots & \\ \cdots & 3 & \cdots & \cdots & \cdots & 2 & \cdots \\ & \vdots & & & & \vdots & \\ & \vdots & & & & \vdots & \\ \cdots & 2 & \cdots & \cdots & \cdots & 2 & \cdots \\ & \vdots & & & & \vdots & \end{pmatrix} \end{array}.$$

We see that the diagonal entries $(PP')_{ii}$ counts the number of locations person i traverses within this time frame, while the off-diagonal entries $(PP')_{ij}$ count the number of locations person i and person j have in common. Note however, that if the time frame is long enough, having a location in common does not guarantee that person i and person j actually cross path. It may well be that person i enters and leaves the said location before person j arrives on scene. For this reason, we do not aggregate the person-by-location matrices into time-frames of more than one second, to ensure that if two people are in the same location within this time-frame, they must be in close enough physical proximity for disease-spread to be viable.

The resulting product PP' is a $n \times n$ matrix, where

$$(PP')_{ij} = \begin{cases} 1 & \text{if person } i \text{ and person } j \text{ were in the same location at time } t \\ 0 & \text{otherwise} \end{cases}$$

with diagonal entries $(PP')_{ii} = 1$ if person i is inside the ED at time t. Because we do not consider an individual to be linked to itself, we remove the non-zero diagonal entries to obtain the (simple) adjacency matrix A:

$$A = PP' - \text{diag}(PP').$$

Doing this for every second $t = 1, \ldots, 43202$, we obtain a sequence of person-by-person adjacency matrices $\{A^{[t]}\}$, which is representative of a dynamic network with nodes as individuals and possible disease-transmitting contacts between them (irrespective of location) represented by edges. The time-evolution of the network is thus captured by the appearance and disappearance of edges over time.

The RFID tags transmitted their unique identifier every 10 seconds [50]. For this reason, the per-second data is considered incomplete and we therefore work with adjacency matrices at the 10-second time resolution. Explicitly, we will divide the 12-hour shift into intervals of 10 seconds. Suppose an interval consists of the times $t_1, \ldots, t_{10}$. Then the aggregated adjacency matrix $A^{[t']}$ over this interval is given by

$$\left(A^{[t']}\right)_{ij} = \begin{cases} 1 & \text{iff } \left(A^{[t_k]}\right)_{ij} = 1 \text{ for some } k \in \{1, \ldots, 10\} \\ 0 & \text{otherwise.} \end{cases}$$

We point out that the total contact time between two people within the 10-second interval is not taken into account. It is possible that two people were in the same location for longer than 1 second, or that they crossed paths at different locations within the 10-second interval.

We proceed with the assumption that the aggregated matrix $A^{[t']}$ has a 1 in the ij-th entry if, at any point within a 10-second time-frame, person i and person j made at least a 1-second contact somewhere in the ED.

2.2 Temporal dynamics of contact network data

2.2.1 Temporal dynamics of degree

The degree of a node refers to the number of edges associated with it. In this dynamic network, node degree evolves over time, and is interpreted as the number of distinct potential disease-spreading contacts per time-step. Given an adjacency matrix A associated with a simple, undirected network, the degree of node i is given by the sum of the entries in ith row (or column) of A. The maximum degree is given by

$$\max_i \sum_j (A)_{ij}$$

which is a matrix norm often written as $\|A\|_1$. The spectral radius of A is defined as

$$\rho(A) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}.$$

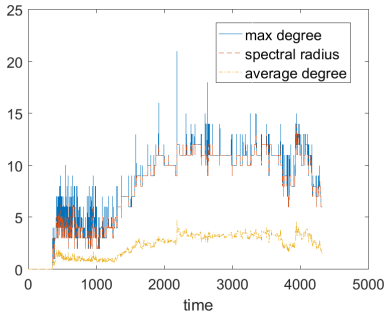
It is well-known that the spectral radius of A is a lower bound for any matrix norm on A, hence

$$\rho(A) \leq \|A\|_1 = \max_i \sum_j (A)_{ij}.$$

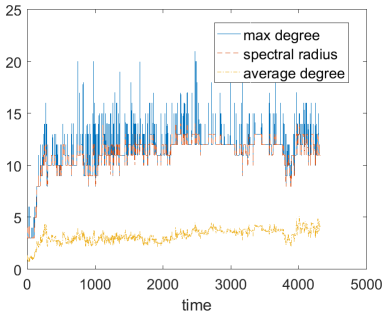
In Figure 2.2 we show the time-evolution of maximum degree, spectral radius and average degree, where the average is computed over the number of people in the ED at that time step. Note that the count of persons in the ED includes people who are not necessarily interacting with others. It is interesting to observe that the spectral radius tracks the

Figure 2.2: Temporal dynamics of maximum and average degree relative to spectral radius. Note that average degree is computed by dividing total degree by the number of people in the ED at that time step, and this count includes people who are not necessarily interacting with others.

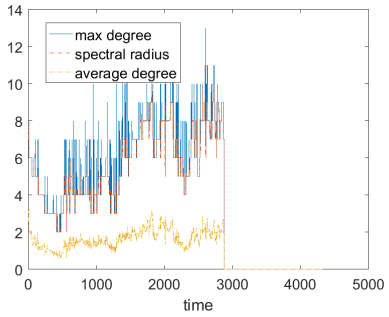
(a) Shift 1



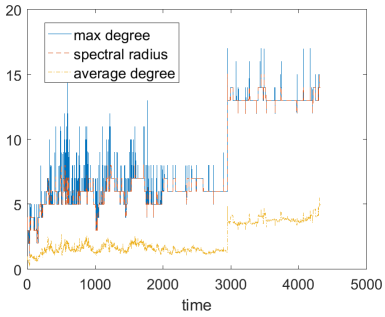
(b) Shift 2



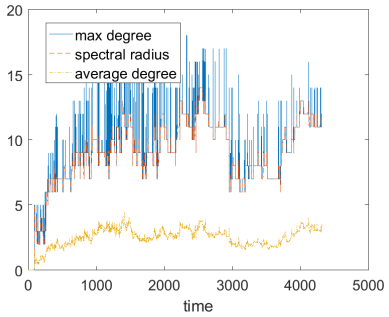
(c) Shift 3



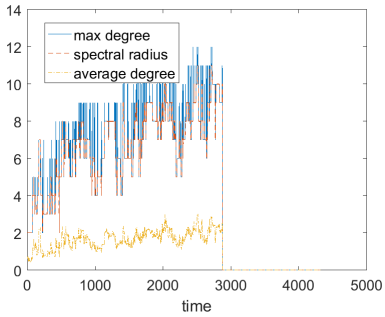
(d) Shift 4



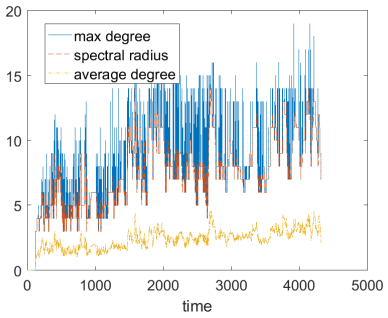
(e) Shift 5



(f) Shift 6



(g) Shift 7



maximum degree per time-step very closely.

2.2.2 Temporal dynamics of pairwise interactions by type: SS, SP, PP

We consider the following interaction types: staff-staff (SS), staff-patient (SP) and patient-patient (PP). (Because the data is undirected, we do not distinguish between staff-patient and patient-staff interactions.) In Figure 2.3 we plot the temporal dynamics of the different types of pairwise interactions, based on the 10-second aggregated data. Observe that consistently across all seven shifts, and irrespective of time, SS interactions dominate SP and PP interactions.

2.3 Locations analysis

Recall that in the construction of the person-by-person adjacency matrices, information about *where* the interactions took place is lost. The network-based analysis is therefore unable to make use of this rich resource of information. We study the location information separately from the network analysis, and present the findings in this section.

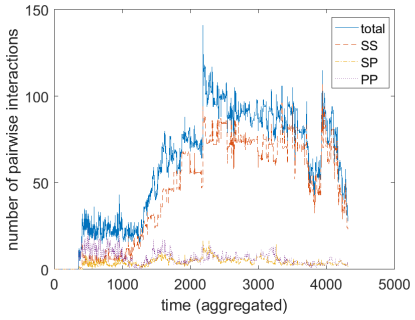
The contact data from the person-by-location matrices $P^{[t]}$ described in Section 2.1 is stored in quadruples of the form $[i, j, \text{location}, \text{time}]$, indicating that nodes i and j shared a location at a certain time-step. Note that when analyzing location information, we work with 1-second time-frames: we do not aggregate into 10-second intervals (as done when constructing the temporal network) because aggregation may result in more than one location associated with each interaction pair.

2.3.1 Locations associated with interaction types: SS, SP, PP

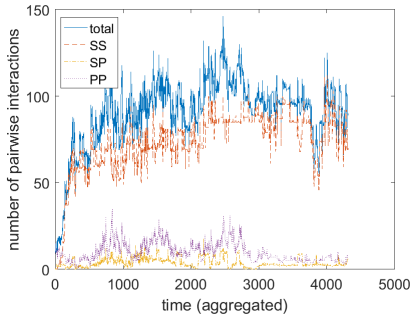
Figure 2.5 shows the distribution of pairwise interactions over all 95 zones in the ED. For visual interpretative ease we plot the square-root of the number of interactions. Note that the square-root distorts the frequencies slightly, since $\sqrt{a+b} \neq \sqrt{a} + \sqrt{b}$, but bearing

Figure 2.3: Temporal dynamics of (aggregated) pairwise interactions stratified by interaction type: SS, SP and PP. In (h) we aggregate the interactions over time.

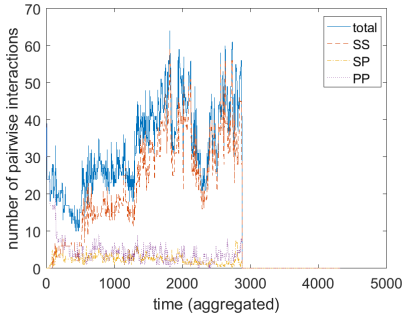
(a) Shift 1



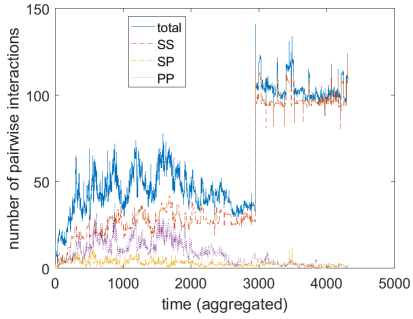
(b) Shift 2



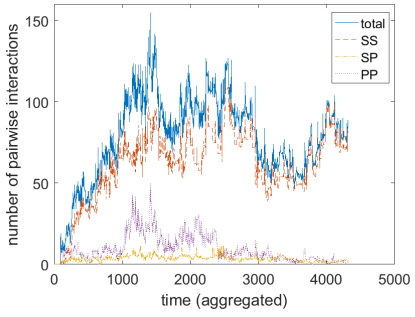
(c) Shift 3



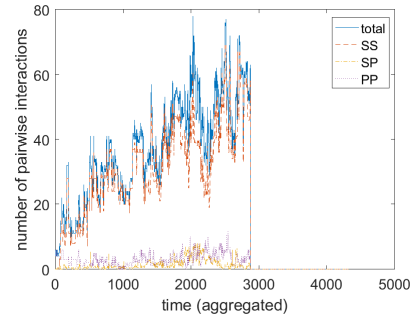
(d) Shift 4



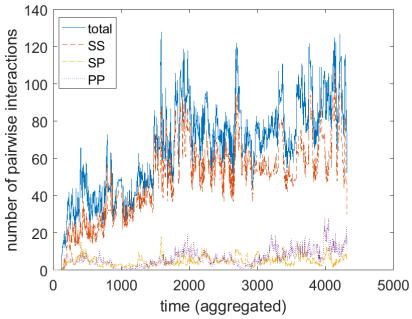
(e) Shift 5



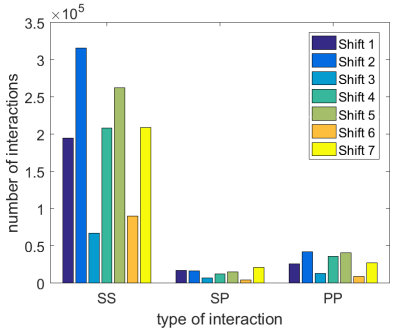
(f) Shift 6



(g) Shift 7



(h) Summary



this in mind, it is nonetheless helpful to have a global picture of how the interactions are distributed across the ED. It is evident that location 88 (triage registration) is by far the busiest across all shifts. Additionally, locations 92 and 94 (both are ED waiting areas) stand out because they are the only locations at which interactions involving patients (SP and PP) consistently outnumber SS interactions.

2.3.2 Most frequented locations

Per interaction type, we do the following: **1)** find the top 5 locations per shift; **2)** for each of the unique locations in this list, rank these locations based on frequency (per shift); **3)** average the rankings over all shifts; **4)** pick the top 5 locations based on the average ranking. The number of interactions at each of these top 5 most frequented locations is shown in Figure 2.6. Descriptions of the corresponding locations are shown in Table 2.2. We conclude that location 88 (triage registration) is the most frequented across all interaction types. Interactions involving patients also occur quite frequently at locations 92 and 94 (ED waiting areas), but with respect to only SS interactions, neither location ranks among the top 5. At both ED waiting areas, PP interactions outnumber SP interactions. This analysis shows that apart from the triage registration area, patient activity occurs mainly in the ED waiting areas, while most of the staff activity takes place in other parts of the ED. We also see clearly that staff interactions amongst themselves are an order of magnitude larger than their interactions with patients.

2.3.3 Inactive locations

A location is called 'inactive' if no interactions occur there over all seven shifts. We show in Figure 2.4 the number of inactive locations per interaction type. This information can potentially be useful in disease-prevention strategies, or future designs of the ED to maximize utility. Further analysis can be performed to address specific questions.

Figure 2.4: Inactive locations per interaction type. These are locations where no interactions took place over all seven shifts.

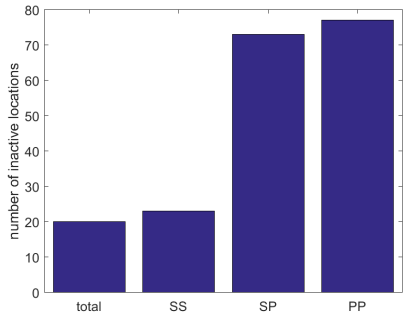
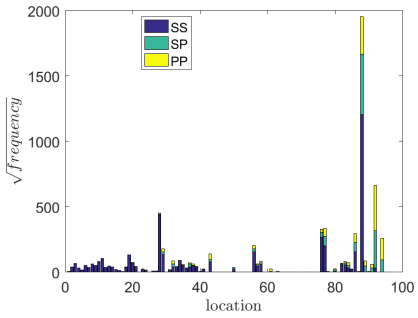
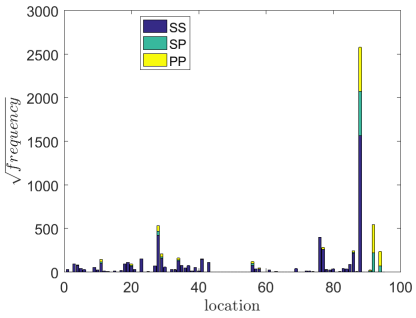


Figure 2.5: Pairwise interactions per location stratified by interaction type: SS, SP and PP.

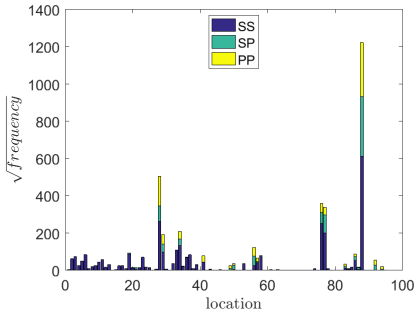
(a) Shift 1



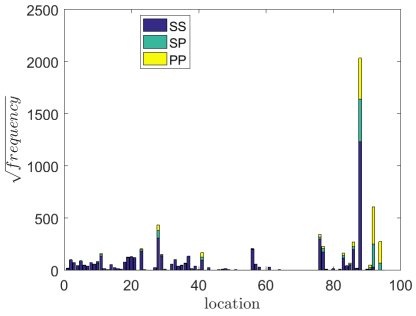
(b) Shift 2



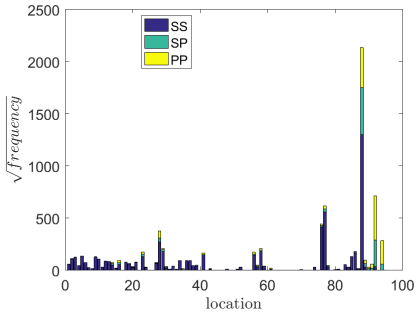
(c) Shift 3



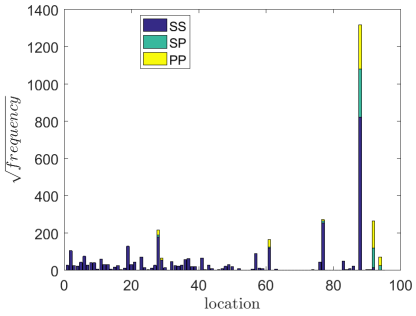
(d) Shift 4



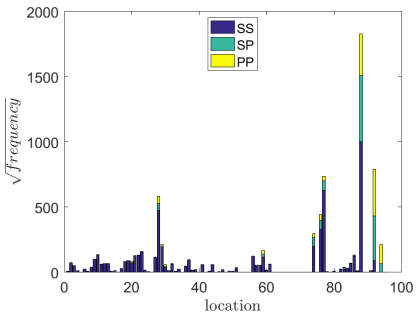
(e) Shift 5



(f) Shift 6



(g) Shift 7



2.3.4 Temporal dynamics of interactions at frequent locations

In Figure 2.7 we plot the time-evolution of interactions at the locations of interest pointed out in Section 2.3.1. Over all shifts, SS interactions dominate at the triage registration area (88). However there are interaction patterns that appear to be shift-specific. For example, in Shift 5, there were more SP interactions at triage registration compared to other shifts, perhaps indicative of an atypical group of patients requiring treatment during that time.

Figure 2.6: Top 5 most frequented locations per interaction type. Refer to Table 2.2 for descriptions of the locations.

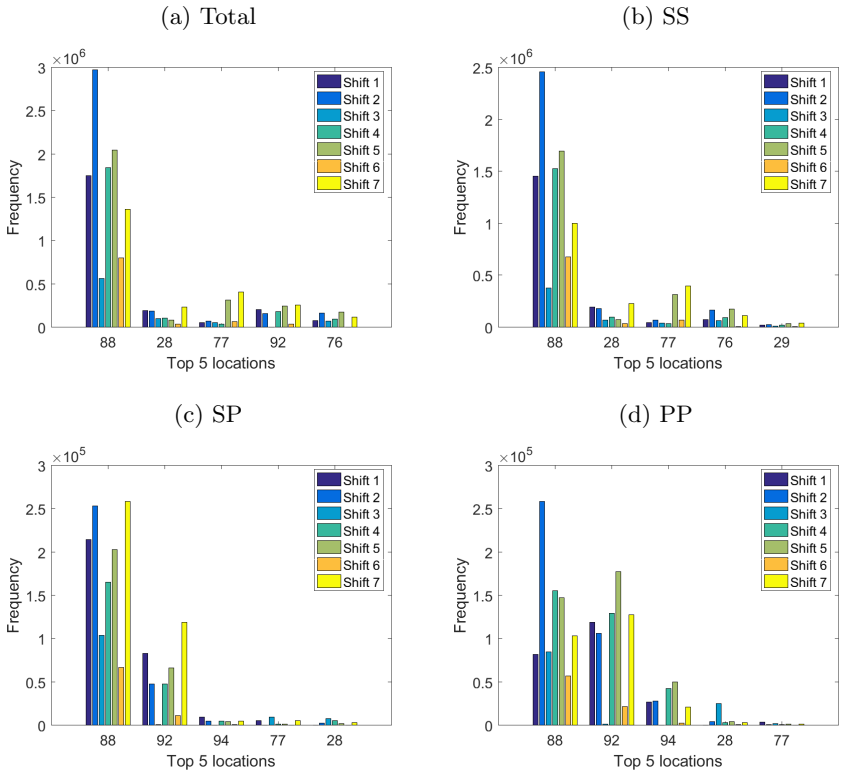
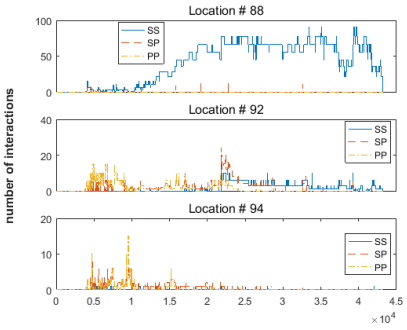


Table 2.2: Descriptions of the top 5 most frequented locations over all interaction types

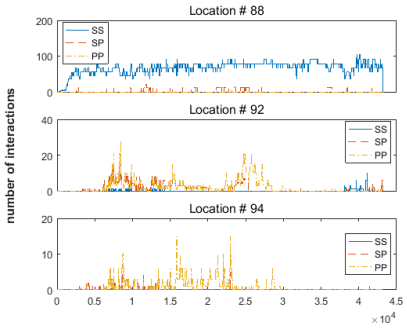
Location	Description	Type	Area (sqft)
28	ED Nurse Station R	Staff Support	1469
29	ED Nurse Station L	Staff Support	1469
76	Office Area	Administrative Support	708
77	Staff Break Area	Administrative Support	695
88	Triage & Registration	Patient Care	282
92	ED Waiting Area	Primary Waiting Area	1888
94	ED Waiting Area	Secondary Waiting Area	1888

Figure 2.7: Temporal dynamics of interactions at frequent locations

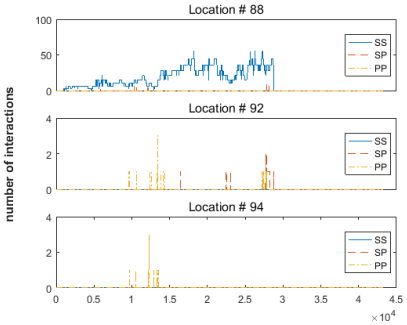
(a) Shift 1



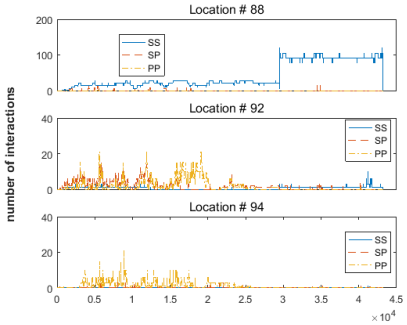
(b) Shift 2



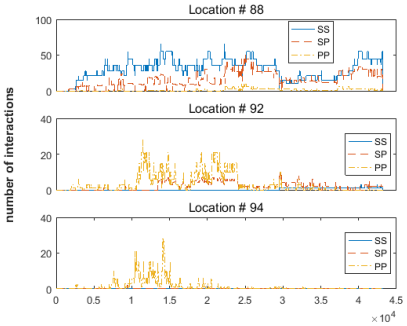
(c) Shift 3



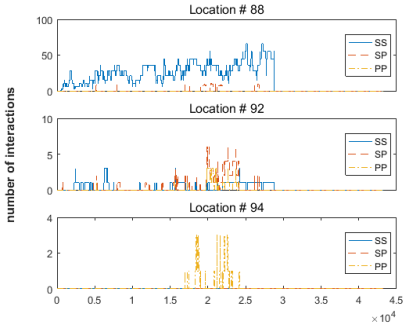
(d) Shift 4



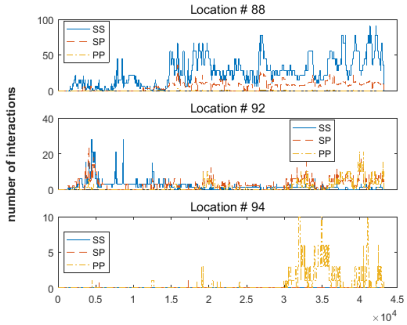
(e) Shift 5



(f) Shift 6



(g) Shift 7



Chapter 3

Walk-based centrality measures

In this chapter, we present the network-based centrality measures that are utilized in this work. Centrality measures are designed to capture some notion of 'importance' based on the underlying structure of the connections between nodes in the network. The resulting numerical values assigned to each node in the network can then be used to *rank* the nodes relative to each other. This is an active area of research, and many different measures have been proposed and studied in depth [18, 26, 59]. Different notions of importance are emphasized by the various measures, and it is often the case that a node may rank highly with respect to one measure, but not with respect to another. Centrality measures are therefore used to provide some insight into the structural relationship between nodes, but must always be interpreted with the underlying notion of importance in mind. The use of different centrality measures on the same network can be instructive and the observed differences can further enhance our understanding of the relationship between nodes in the network.

We are motivated by examples in [52] and [75], which show that centrality measures based on time-aggregated versions of temporal networks (as shown in Fig. 1.2) fail to adequately capture important nodes. The work in [65] suggests that temporal correlations in network

data should not be underestimated and that consequently, detecting important individuals based on temporal structure may have a significant impact on targeted intervention strategies. In [69] it was shown that while centrality measures on the time-aggregated network (such as node degree, betweenness centrality, eigenvector centrality) can improve immunization strategies, node strength, which measures the total time exposed to others, was the most effective. These results illustrate the importance of utilizing temporal information in the context of identifying key players in the contagion process.

Our work aims to bring recently developed temporal centrality measures into the analysis. Temporal centrality measures provide a means of identifying important nodes based solely on time-dependent network structure. While temporal centrality is a relatively young field compared to its static counterpart, new measures are constantly being developed [39, 44, 75, 76]. We focus our attention on *dynamic communicability* introduced by Grindrod *et al.* [37]. This is a generalization of Katz centrality [41] to the temporal setting. To set the stage, we will first discuss Katz centrality in Section 3.1. The generalization to dynamic communicability is described in Section 3.2, and in Section 3.3 we discuss the class of walk-based centrality measures in which dynamic communicability belongs. We will show that walk-based centrality measures can be expressed neatly as matrix functions on the underlying adjacency matrices, which has the advantage of exploiting well-developed numerical linear algebraic techniques for efficient and accurate computation. We emphasize that the computation of centrality scores depends solely on the network structure and is completely independent of any contagion process on the network.

3.1 Katz centrality

Consider a static, directed network on n nodes, with an associated $n \times n$ binary adjacency matrix A, where $A_{ij} = 1$ if and only if there is an edge between node i and node j. We allow for $A_{ij} \neq A_{ji}$ so that the adjacency matrix A may be non-symmetric. We will assume

that there are no self-loops or multiple edges between nodes – such networks are sometimes referred to as 'simple'. While traditional Katz centrality [41] is typically computed on an undirected network (where the underlying adjacency matrix A is symmetric), here we discuss Katz's method in generality where the edges of the network are associated with a direction.

A walk on the network is a sequence of edges connecting nodes, where both nodes and edges can be repeated. More formally, we define a walk on the network as follows:

Definition 3.1.1. *A walk of length w from node v_0 to node v_w consists of a sequence of nodes and edges $v_0, e_1, v_1, e_2, \ldots, e_w, v_w$, such that for $1 \leq i \leq w$, the edge e_i has endpoints v_{i-1} and v_i. Equivalently, a walk may be specified as a sequence of nodes $v_0, \ldots, v_w$ where for $0 \leq i \leq w-1$, the edge (v_i, v_{i+1}) is present in the network.*

A walk is said to be closed if $v_0 = v_w$. We emphasize that in a walk as defined, it is possible for $e_i = e_j$ where $i \neq j$. That is to say, the same edge can be traversed multiple times in a walk. This is in contrast to a *path*, where neither nodes nor edges can be reused.

Walks between nodes can be considered as means for communication or interaction. We also place more emphasis on short walks as these are less prone to noise and can therefore arguably transmit information more accurately. To quantify node i's ability to communicate with node j, we consider walks of all possible lengths between i and j, where longer walks are proportionally downweighted. We will see that the resolvent of the matrix A captures this information in a very natural way.

The resolvent of a matrix A is defined as $(I - \alpha A)^{-1}$, which exists provided $1/\alpha \notin \sigma(A)$, where $\sigma(A)$ denotes the spectrum of A. If in addition, α satisfies the condition $0 < \alpha < 1/\rho(A)$, where $\rho(A)$ is the spectral radius of A, the power series $I + \alpha A + \alpha^2 A^2 + \ldots$ converges to the resolvent $(I - \alpha A)^{-1}$. This can easily be seen by multiplying the power series with

$(I - \alpha A)$. We can therefore identify the resolvent with its power series representation

$$(I - \alpha A)^{-1} = I + \alpha A + \alpha^2 A^2 + \ldots + \alpha^w A^w + \ldots$$

provided $0 < \alpha < 1/\rho(A)$.

It is a well-known graph-theoretic fact that $(A^w)_{ij}$ counts the number of walks of length w beginning at node i and ending at node j. The ijth entry of $(I - \alpha A)^{-1}$ is therefore a measure of the number of walks starting at node i and ending at node j, where walks of length w are downweighted by α^w. Summing over all nodes j, we obtain a measure of how well node i communicates, or *broadcasts* information to the network as a whole. Explicitly, the broadcast centrality of node i is given by

$$\begin{aligned}
\left[(I - \alpha A)^{-1} \cdot \mathbb{1}\right]_i &= \sum_{j=1}^{n} \left[(I - \alpha A)^{-1}\right]_{ij} \\
&= \sum_{j=1}^{n} \left[I + \alpha A + \alpha^2 A^2 + \ldots\right]_{ij} \\
&= 1 + \alpha \left(A \cdot \mathbb{1}\right)_i + \alpha^2 \left(A^2 \cdot \mathbb{1}\right)_i + \ldots
\end{aligned}$$

where $\mathbb{1}$ is the vector of all ones, and $\left(A^w \cdot \mathbb{1}\right)_i$ counts all possible walks of length w *beginning* at i. Analogously, the jth entry of the column sums of $(I - \alpha A)^{-1}$ can be broken down into a power series of terms of the form $\sum_{i=1}^{n} (A^w)_{ij} = \left(A^w \cdot \mathbb{1}\right)_j$, each counting walks of length w *ending* at node j, therefore providing a measure of node j's ability to *receive* information from the rest of the network.

In a directed network, the dual notions of broadcasting and receiving are given by the row and column sums of the resolvent $(I - \alpha A)^{-1}$ respectively, for a suitable choice of the parameter α. Clearly if the underlying network is undirected, as in the typical application of Katz centrality, the corresponding adjacency matrix A is symmetric and there would be no distinction between row and column sums of either A or powers of A, and hence also

of the resolvent $(I - \alpha A)^{-1}$. In other words, in an undirected network, since there is no distinction between being at the starting or end point of a walk, there is correspondingly no distinction between a node's ability to either *broadcast* or *receive* information.

Observe that the bounds on α also force $(I - \alpha A)^{-1}$ to be non-negative, as $I - \alpha A$ is a non-singular M-matrix. Hence, the row/column sums are positive and can thus be used for ranking purposes. Heuristically, a node that is highly ranked according to Katz centrality is one which is connected via 'short' walks to 'many' other nodes in the network. Note that the terms 'short' and 'many' are relative: a node with the same number of walks to the same number of nodes could have a very different ranking in another network. It is important to bear in mind that the ranking of a node is intrinsically tied to the structure of the network as a whole and therefore on the attributes of the other nodes in the network.

3.2 Dynamic communicability

Dynamic communicability is a centrality measure introduced by Grindrod, Parsons, Higham and Estrada in 2011 [37], which generalizes Katz centrality to the dynamic setting. To see this, we need to consider dynamic, or temporal, walks. Consider now a time-evolving network on n nodes, which is represented by a sequence of adjacency matrices $A^{[t]}$ for $t = 1, \ldots, M$, where t indexes the time-step. Each matrix $A^{[t]}$ is symmetric, so that the network at each time-step is undirected.

Definition 3.2.1. *A dynamic, or temporal, walk of length w from node v_0 to node v_w consists of a sequence of nodes $v_0, \ldots, v_w$ and non-decreasing time-indices $t_1 \leq t_2 \leq \ldots \leq t_w$ where for $0 \leq i \leq w - 1$,*

$$\left(A^{[t_i]}\right)_{v_{i-1}, v_i} = 1.$$

In other words, a dynamic walk is a sequence of edges connecting nodes, with the added requirement that the subsequent edge must come from either the same time-step as the

preceding edge, or any time-step thereafter. Note that we do not require the edges to come from either consecutive, or completely distinct time-steps. The only requirement is for the time-indices of the edges to be *non-decreasing*. This is imperative, as, in the words of [37], it is this condition which ensures that the walk respects 'the arrow of time' and is therefore dynamic in a physically legitimate sense.

Recall that in the static case (where $M = 1$), $\left(A^w\right)_{ij}$ counts the number of walks of length w beginning at node i and ending at node j. Analogously, it is easy to see that

$$\left(A^{[t_1]}A^{[t_2]}\dots A^{[t_w]}\right)_{ij}$$

counts the number of dynamic walks between node i and j, where the kth edge in the walk comes from $A^{[t_k]}$. Note that even though the base matrices are symmetric, the product $A^{[t_1]}A^{[t_2]}\dots A^{[t_w]}$ is not. In physical terms, this translates to the fact that a dynamic walk from i to j does not guarantee a dynamic walk from j to i.

Analogous to Katz centrality, we will use the matrix resolvents to succinctly summarize dynamic walk information in the following way. For the sake of argument, consider the case where $M = 2$ and let α be an appropriate constant. Then

$$\left(I + \alpha A^{[1]} + \alpha^2\left(A^{[1]}\right)^2 + \dots\right)\left(I + \alpha A^{[2]} + \alpha^2\left(A^{[2]}\right)^2 + \dots\right)$$

has the following terms containing α^3

$$\alpha^3\left(A^{[1]}\right)^3 + \alpha^3\left(A^{[2]}\right)^3 + \alpha^3\left(A^{[1]}\right)^2 A^{[2]} + \alpha^3 A^{[1]}\left(A^{[2]}\right)^2$$

and these contain all the information about dynamic walks of length 3. The first term counts walks where all three edges occur at the first time-step; the second term counts walks where all three edges occur at the second time-step; the third term counts walks where the first

two edges occur at the first time-step and the third edge occurs at the second time-step; the last term counts walks where the first edge occurs at the first time-step and the next two edges occur at the second time-step. Therefore, for an arbitrary number of time-steps $M \geq 2$, by expressing the resolvent of each matrix as a power series, we see that all products of the form

$$\alpha^w A^{[t_1]} A^{[t_2]} \dots A^{[t_w]}$$

where $w \geq 1$ and $t_1 \leq t_2 \leq \ldots \leq t_w$ are contained in the product

$$Q := \left(I - \alpha A^{[1]}\right)^{-1} \left(I - \alpha A^{[2]}\right)^{-1} \cdots \left(I - \alpha A^{[M]}\right)^{-1}. \tag{3.1}$$

Q is called the *dynamic communicability matrix*. For the power series representation to converge, we require that for $1 \leq t \leq M$,

$$0 < \alpha < \frac{1}{\max\limits_t \rho(A^{[t]})}.$$

Observe that all dynamic walk information is now summarized into a single matrix Q. Q_{ij} can be interpreted as a weighted count of dynamic walks of all possible lengths from node i to node j, where walks of length w are downweighted by the factor α^w. This is a measure of 'dynamic communicability' between node i and node j, with node i being the broadcaster and node j being the receiver. Summing over all receivers j, we obtain

$$\sum_{j=1}^{n} Q_{ij} = \left(Q \cdot \mathbb{1}\right)_i$$

which is a measure of how well node i broadcasts information to the rest of the network as a whole. The row sums therefore provide a notion of *broadcast centrality*. On the other

hand,

$$\sum_{i=1}^{n} Q_{ij} = \left(Q^T \cdot \mathbb{1} \right)_j$$

is a measure of how well node j receives information from all other nodes in the network, and therefore provides a notion of *receive centrality*.

Definition 3.2.2. *Given the dynamic communicability matrix as defined in Eq. (3.1), the vector of broadcast centralities, denoted BC, is given by*

$$BC = Q \cdot \mathbb{1},$$

while the vector of receive centralities, denoted RC, is given by

$$RC = Q^T \cdot \mathbb{1}.$$

The matrix Q is therefore able to capture dual notions of broadcasting and receiving. We reiterate that Q is in general not symmetric, therefore the row and column sums are in general different from each other. We also point out that because the base matrices $A^{[t]}$ are symmetric, the resolvents $(I - \alpha A^{[t]})^{-1}$ are also symmetric. Furthermore, since for any square matrix A, $\left(A^{-1}\right)^T = \left(A^T\right)^{-1}$, we have

$$\begin{aligned} Q^T &= \left(\left(I - \alpha A^{[1]}\right)^{-1} \left(I - \alpha A^{[2]}\right)^{-1} \cdots \left(I - \alpha A^{[M]}\right)^{-1} \right)^T \\ &= \left(I - \alpha A^{[M]}\right)^{-1} \left(I - \alpha A^{[M-1]}\right)^{-1} \cdots \left(I - \alpha A^{[1]}\right)^{-1}. \end{aligned} \tag{3.2}$$

In other words, broadcast and receive centralities are related by a reversal of the time ordering.

Synthetic examples in [37] and [52] illustrate that BC and RC measures perform better than aggregated measures in identifying nodes with time-sensitive links as important. The

term *dynamic communicator* coined in [52] refers precisely to the nodes which rank highly in the dynamic sense but do not stand out in a snapshot or aggregate view of the network.

3.2.1 Limit as $\alpha \to 0$

Consider a two-term product of the form

$$\begin{aligned}
&\left(I - \alpha A^{[1]}\right)^{-1} \left(I - \alpha A^{[2]}\right)^{-1} \\
=&\left(I + \alpha A^{[1]} + \alpha^2 \left(A^{[1]}\right)^2 + \alpha^3 \left(A^{[1]}\right)^3 + \mathcal{O}(\alpha^4)\right) \left(I + \alpha A^{[2]} + \alpha^2 \left(A^{[2]}\right)^2 + \alpha^3 \left(A^{[2]}\right)^3 + \mathcal{O}(\alpha^4)\right) \\
=&I + \alpha \left(A^{[1]} + A^{[2]}\right) + \alpha^2 \left(\left(A^{[1]}\right)^2 + \left(A^{[2]}\right)^2 + \underbrace{A^{[1]} A^{[2]}}_{\text{asymmetry}}\right) + \mathcal{O}(\alpha^3).
\end{aligned}$$

Generalizing to products with M terms, we can write the communicability matrix Q as

$$Q = \left(I - \alpha A^{[1]}\right)^{-1} \ldots \left(I - \alpha A^{[M]}\right)^{-1} = I + \alpha \left(\sum_{t=1}^{M} A^{[t]}\right) + \mathcal{O}(\alpha^2)$$

and so

$$BC = Q \cdot \mathbb{1} = I \cdot \mathbb{1} + \alpha \left(\sum_{t=1}^{M} A^{[t]}\right) \cdot \mathbb{1} + \mathcal{O}(\alpha^2).$$

Shifting by $\mathbb{1}$ and scaling by the constant α, we have

$$\frac{BC - \mathbb{1}}{\alpha} = \frac{Q \cdot \mathbb{1} - \mathbb{1}}{\alpha} = \underbrace{\left(\sum_{t=1}^{M} A^{[t]}\right) \cdot \mathbb{1}}_{AD} + \mathcal{O}(\alpha). \tag{3.3}$$

Observe that the first term on the RHS ranks nodes based on the degree of the aggregated matrix

$$\sum_{t=1}^{M} A^{[t]}$$

which we term aggregate degree, abbreviated by AD. Since a shift and scale of the BC vector will not change the associated BC rankings, Eq (3.3) means that as $\alpha \to 0$, BC

rankings should theoretically converge to AD rankings. A similar argument shows that the same is true for RC rankings. We will use this fact as a test for the numerical accuracy of our results, as well as to inform our choice for the parameter α. Recall that α must be chosen so that

$$0 < \alpha < \frac{1}{\max_t \rho(A^{[t]})}.$$

We aim to choose α sufficiently far away from 0 so that we do not merely replicate aggregate degree. On the other hand α cannot be too close to upper limit, as in this case, the matrix $I - \alpha A^{[t]}$ for which the maximum spectral radius is attained will be close to singular, and the entries of its inverse will dominate the computation of Q. In this regime, the computation of Q will be sensitive to small changes in α.

3.2.2 Possible modifications

As discussed in [37], in order to eliminate the possibility of long walks, or closed walks such as $i \mapsto j \mapsto i$ taking place within a single time-step, we may enforce the walks to use at most one link per time window by using this modified version of Q where each term in the product is a first-order approximation of the matrix resolvent:

$$Q = (I + \alpha A^{[1]})(I + \alpha A^{[2]}) \dots (I + \alpha A^{[M]}).$$

This contains all products of the form $\alpha^w A^{[t_1]} \dots A^{[t_w]}$ where $t_1 < \dots < t_w$ are all distinct. This modification may be appropriate in ultra-high frequency regimes, where the length of each time window is so small that it is physically unfeasible for information to pass through more than one edge per time step.

It is reasonable to think that walks that started recently are more important than walks that started long ago. Based on this motivation, Grindrod and Higham [35] developed a formulation which controls for the history of walks. That is, in addition to downweighting

walks based on length, walks are also downweighted based on time of origin. The formulation is based on a matrix iteration of the form

$$S^{[k]} = (I + e^{-b\Delta t_k} S^{[k-1]})(I - \alpha A)^{-1}.$$

Most recently, the same authors present a dynamical systems view of dynamic communicability [36] which generalizes to continuous time:

$$S(t + \delta t) = (I + e^{-b\delta t} S(t))(I - \alpha A(t + \delta t))^{-\delta t} - I.$$

3.2.3 Data studied

Dynamic communicability has been studied on telecommunication data (MIT [37]) and email data (Enron [37], [52]). It has also been used to characterize learning in the human brain [51]. We would like to study BC/RC measures in the context of disease spread on a person-to-person contact network. In particular, we want to see how a seed node's BC ranking/measure is related to epidemic outcome. We point out that [53] studies the same measure, dynamic communicability, in relation to contagion on a temporal network. Our work differs in the methodology of the infection simulation, as well as in the use of regression to analyze the resulting simulation output. Additionally, the network studied in [53] is based on email communication, on which the notion of epidemic spread is inherently different from that of a proximity-based contact network. In [60], dynamic communicability was used to find a subset of nodes to maximize influence on the SocioPatterns hospital ward dynamic contact network. Our work does not focus only on the highly central nodes, but aims to evaluate the overall effect of dynamic communicability on epidemic outcome.

3.2.4 Relationship with the matrix exponential

We point out that the matrix exponential

$$e^A = I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \dots$$

provides another walk-based centrality measure which is in principal identical to (static) Katz centrality described in 3.1. In the temporal setting, one can use the matrix exponential instead of the matrix resolvent to compute a measure of dynamic communicability [27]. The rationale of downweighting long walks remains, with the difference lying in the downweighting factors themselves: $1/w!$ versus α^w. By replacing the matrix resolvents with matrix exponentials, we obtain another walk-based measure in the following way:

$$\tilde{Q} = e^{A^{[1]}} e^{A^{[2]}} \cdots e^{A^{[M]}}. \tag{3.4}$$

One difficulty that arises with using the matrix exponential is that the downweighting factor of $1/w!$ typically penalizes walks of length w much more severely than α^w. In the static case where there is only one adjacency matrix A, this property can be useful in revealing more information about mid-ranking nodes the network [34]. However, in the dynamic case, when there are many adjacency matrices to work with, the severe downweighting of long walks may lead to an inability to distinguish between nodes. Preliminary experiments on our data set are indicative of this: Normalization during the computation of $\tilde{Q}$ (see Section 4.1.1) results in the majority of nodes with broadcast centrality close to zero, and only two nodes with non-zero values (see Section 4.2). On the other hand, dynamic communicability based on the matrix resolvent was able to clearly distinguish about 20 nodes which had broadcast centrality significantly different from zero.

A possible remedy is to include a tuning parameter β in the matrix exponential as follows:

$$e^{\beta A} = I + \beta A + \frac{\beta^2}{2!}A^2 + \dots .$$

The parameter β can be viewed, from a thermodynamical point of view, as a form of inverse temperature of the system [27]. Since β has no upper bound (unlike α in dynamic communicability), the factor $\beta^k/k!$ can potentially give longer walks more weight than α^k. It may be of interest to perform a comparative study of the results obtained by both methods. We will leave this for future work: in the rest of this paper, we discuss results based on the resolvent-based formulation as expressed in Eq. (3.1).

Reference [37] explains the relationship between the dynamic communicability matrix Q and the matrix exponential in the following way: Suppose that there is a physical constraint limiting the length of time during which information can be collected. If we narrow the time windows beyond this point, so that we end up with a contiguous collection of K identical adjacency matrices A, there arises, in the computation of Q, a term of the form

$$(I - \alpha A)^{-K} .$$

In this limit, suppose we allow the downweighting parameter α to scale inversely with the length of the time window. Then we see that Q contains a quantity of the form

$$\lim_{K\to\infty} \left(I - \frac{\alpha}{K}A\right)^{-K} = \exp(\alpha A).$$

We see therefore that the matrix exponential arises in this way only when the length of the time window is shorter than the time required to capture new information.

3.3 Classes of centrality measures

We have seen in Section 3.2.4 that the matrix resolvent $(I-\alpha A)^{-1}$ and the matrix exponential $e^{\beta A}$ are two matrix functions that succinctly capture walk-based information on static networks, and the same principles generalize naturally to the temporal setting.

In the static case, where there is only one adjacency matrix A, much work has been developed on centrality measures based on these walk-based interpretations of matrix functions. E. Estrada developed the concept of subgraph centrality [28] which is based on the diagonal entries of e^A. The trace of e^A is sometimes referred to as the Estrada Index [23]. The row/-column sums of e^A were introduced by M. Benzi and C. Klymko in [15] as a measure called total communicability. In [4], F. Arrigo and M. Benzi used the resolvent formulation (i.e., Katz centrality) on a directed network to obtain the notions of broadcasting and receiving.

A natural question that needs to be addressed is the choice of the parameters α and β. It has been shown by M. Benzi and C. Klymko in [16] that as α approaches 0, Katz-based node rankings converge to degree-based rankings. On the other hand, as α approaches the upper limit of $1/\rho(A)$, Katz-based rankings converge to rankings produced by eigenvector centrality. A similar result is shown for node rankings based on total communicability. The relationship between these various centrality measures is shown heuristically in the following way:

Table 3.1: Relationship between various walk-based measures

	$\alpha \to 0$		$\alpha \to 1/\sigma(A)$	
degree	$\longleftarrow$	$\left[(I-\alpha A)^{-1}\cdot \mathbb{1}\right]_i$ $\left[e^{\beta A}\cdot \mathbb{1}\right]_i$	$\longrightarrow$	eigenvector
	$\beta \to 0$		$\beta \to \infty$	

Eigenvector centrality of a node i can be interpreted as the proportion of infinite walks beginning at node i relative to walks beginning at all other nodes in the network (see [26] p 127 and [22]), and can thus be considered a measure of node i's global influence. We can

therefore think of α as a tuning parameter between local (degree) versus global (eigenvector) influence.

These results on static networks lead to a natural classification of centrality measures based on *walks* versus centrality measures based on *paths*, where the distinction lies in the fact that nodes and edges can be repeated in a walk, but not in path. The relationships described in Table 3.1 allow us to consider the aforementioned centrality measures such as degree, Katz, eigenvector, total communicability and subgraph centrality, as belonging to the class of walk-based centrality. This is in contrast to path-based measures such as betweenness and closeness centrality ([59]).

Since a walk is allowed to re-use the same edge an arbitrary number of times, an edge between two nodes counts as a closed walk of length $2, 4, 6, \ldots$, *ad infinitum*. However, since longer walks are naturally penalized, such pathological realizations do not appear to have much effect in practice. The convenient walk-based expressions based on matrix functions also present computational elegance and ease not available to path-based centrality measures, which are typically computationally intensive, and particularly so in the temporal setting [44]. It is also arguable that contagion on a network does not necessarily follow shortest paths. Diseases, in particular, spread by contact and it is conceivable that in this context, walks form a more realistic trajectory than shortest-paths. While it has been shown, in the static case, that betweenness centrality correlates the strongest with epidemic outcome ([31, 69, 71]) compared to degree and eigenvector centrality, other walk-based measures are not often used in the context of contagion processes. Because the temporal analogue of betweenness centrality is prohibitively expensive to compute [17, 44], this work is not a comparative study between temporal walk-based centrality and temporal path-based centrality. We do not claim that dynamic communicability outperforms temporal betweenness centrality; rather, this work aims to showcase these techniques based on matrix functions to network scientists in other fields who may not be fully cognizant of these

methods developed by mathematicians.

Recall that aggregate degree, AD, ranks nodes according to the degree of the aggregated matrix $\sum_{t=1}^{M} A^{[t]}$. We can binarize the aggregated matrix by replacing any non-zero entry by 1: the resulting binarized matrix has ij-th entry equal to 1 if and only if there exists at least one time step during which nodes i and j were connected. The degree based on the binarized matrix, denoted BD (for binarized degree), therefore ranks nodes solely on the number of distinct contacts made over time, while disregarding all temporal information. Additionally, degree can be viewed as a walk-based measure, where only walks of length 1 are considered. In contrast, AD takes into consideration the *duration* of pairwise interactions in the ranking procedure, but ignores their temporal ordering. Dynamic communicability, on the other hand, relies heavily on the temporal ordering of edges to determine the importance of a node. To do so, the *product* of matrix functions is used to determine centrality, and the discussion in Section 3.2.1 shows that the analogous left-hand limit of Table 3.1 in the temporal setting is AD, suggesting that BC and RC measures based on dynamic communicability can be viewed as more nuanced versions of AD, taking into account walks of length > 1. (There is no obvious analogy to the right hand limit of Table 3.1.) These measures, BD, AD and BC/RC, can therefore be viewed as walk-based centrality measures in increasing temporal complexity, with BD containing no temporal information, AD containing only information regarding the total duration of contacts between nodes, and BC/RC containing the most nuanced temporal information based on the temporal ordering of contacts. The question we seek to answer is this: does the increasing complexity of these measures add value to our understanding of how contagion spreads on the network?

Chapter 4

Dynamic communicability applied to the data

In this chapter we present the results of dynamic communicability applied on the temporal network described in Chapter 2. We work with the contact data of Shift 1 (see Table 2.1). A total of 107 participants agreed to take part in the study, out of which 33 were staff and 74 were patients. An additional 24 patients were in the ED during that time but did not participate in the study.

4.1 Robustness in the choice of α

Recall that the computation of dynamic communicability as defined in Eq. (3.1) requires a choice of the parameter α, where

$$0 < \alpha < \frac{1}{\max\limits_{t} \rho(A^{[t]})} = \alpha_{\max}.$$

For the data based on Shift 1, $\alpha_{\max} = 0.072$. In other words, $\max_k \rho\left(A^{[t]}\right) \approx 14$ is a lower bound for the maximum degree over all time steps, as illustrated in Figure 2.2. In this

section we present the results obtained for the following choices of α:

$$\alpha_1 = 0.25 * \alpha_{\max}$$

$$\alpha_2 = 0.50 * \alpha_{\max}$$

$$\alpha_3 = 0.75 * \alpha_{\max}$$

$$\alpha_4 = 0.85 * \alpha_{\max}.$$

4.1.1 Computational Note

Broadcast and receive centrality scores based on dynamic communicability (as defined in Def. 3.2.2) can be computed in two ways:

- Method I: Compute Q explicitly, for example, as suggested in [37], using an iteration of the form

$$\hat{Q}^{[t]} = \frac{\hat{Q}^{[t-1]} \left(I - \alpha A^{[t]}\right)^{-1}}{\|\hat{Q}^{[t-1]} \left(I - \alpha A^{[t]}\right)^{-1} \|}, \qquad t = 1, 2, \ldots, M,$$

  where $Q^{[0]}$ is the identity matrix, then compute the row and column sums to obtain BC and RC measures.

- Method II: Compute BC and RC measures directly using an iteration of the form

$$BC^{[t]} = \frac{\left(I - \alpha A^{M+1-t}\right)^{-1} \cdot BC^{[t-1]}}{\| \left(I - \alpha A^{M+1-t}\right)^{-1} \cdot BC^{[t-1]} \|}, \qquad t = 1, 2, \ldots, M,$$

  where $BC^{[0]} = \mathbb{1}$. A similar form (using the transpose of the resolvents) is used to compute RC measures.

In both methods, normalization is used to avoid under- or overflow in the computations. Here we use the Euclidean 2-norm, although any matrix norm is applicable. Note that while

normalization changes the absolute values of the centrality measures, it does not change the overall rankings of the nodes.

Table 4.1: Computation times for Method I and Method II

choice of α	Method I (in sec)	Method II (in sec)
$0.25 * \alpha_{\max}$	8.20	0.70 ($\times 2 = 1.40$)
$0.50 * \alpha_{\max}$	17.88	0.79 ($\times 2 = 1.58$)
$0.75 * \alpha_{\max}$	9.83	0.71 ($\times 2 = 1.42$)
$0.85 * \alpha_{\max}$	9.54	0.74 ($\times 2 = 1.48$)

Computations were done using Matlab. For Method I we use `mrdivide` which solves systems of linear equations of the form $\mathbf{x}A = B$; for Method II we use `backslash` which solves a linear system of the form $A\mathbf{x} = \mathbf{b}$. Table 4.1 shows that Method II is an order of magnitude faster than Method I. However, because the first iteration of Method II uses the resolvent of the *last* adjacency matrix $A^{[M]}$, in applications where one needs to compute the BC and RC rankings dynamically, Method I is more appropriate.

It is interesting to note that the computation times depend on α. We do not have an explanation for this behavior.

Both methods yield quantitatively similar results according to multiple measures, as shown in Table 4.2. By manual inspection we noticed that the differences in BC rankings occurred at the low-ranked nodes, while the differences in RC rankings for $\alpha = 0.5 * \alpha_{\max}$ occurred in the second third of the lists, which explains the low Kendall correlation.

In the following sections, where we study the BC/RC measures computed at the end of the shift, we present the results based on the faster method, Method II. In Section 4.4, temporal dynamics of the rankings are computed using Method I.

Table 4.2: Comparisons of results using Method I and Method II

different values of α	0.25	0.50	0.75	0.85
corr(BC)	1.0000	1.0000	1.0000	1.0000
corr(BCranks)	1.0000	1.0000	0.9968	0.9979
kcorr(BCnodes)	1.0000	1.0000	0.8286	0.8551
isim(BCnodes)	0.0000	0.0000	0.0049	0.0039
Top 10 intersection (BC)	10	10	10	10
Top 5 intersection (BC)	5	5	5	5
corr(RC)	1.0000	1.0000	1.0000	1.0000
corr(RCranks)	1.0000	0.9180	0.9999	1.0000
kcorr(RCnodes)	1.0000	0.5909	0.9496	0.9806
isim(RCnodes)	0.0000	0.0483	0.0223	0.0031
Top 10 intersection (RC)	10	10	10	10
Top 5 intersection (RC)	5	5	5	5

4.1.2 BC and RC measures

For interpretative ease, we point out that nodes labeled 1-33 are staff, and nodes labeled 34-107 are patients. We emphasize that the role of centrality metrics is first and foremost to provide a means to rank nodes relative to each other; the numerical values themselves may not be directly interpretable.

From Figure 4.1 we see that most BC and RC measures are close to zero, irrespective of α. Although BC and RC measures cannot be exactly zero (if so, Q must have an entire row or column of zeros which is impossible since Q is an invertible matrix), normalization at each iteration in the computation of Q is likely to result in very small values.

The RC measures exhibit a curious feature: regardless of α, the same seven nodes have non-negligible RC measure, and they all have the same magnitude, agreeing to many decimal places. We point out that these nodes are registered nurses (RN). Since RN's are typically the last people that patients see before leaving the ED, it is conceivable that RN's are often at the receiving ends of walks, which explains why the method ranks them as high receivers.

Figure 4.1: BC and RC measures (Shift 1) associated with different values of α. There is one data point per node; the horizontal axis is the node ID label. Nodes labeled 1-33 are staff and nodes labeled 34-107 are patients.

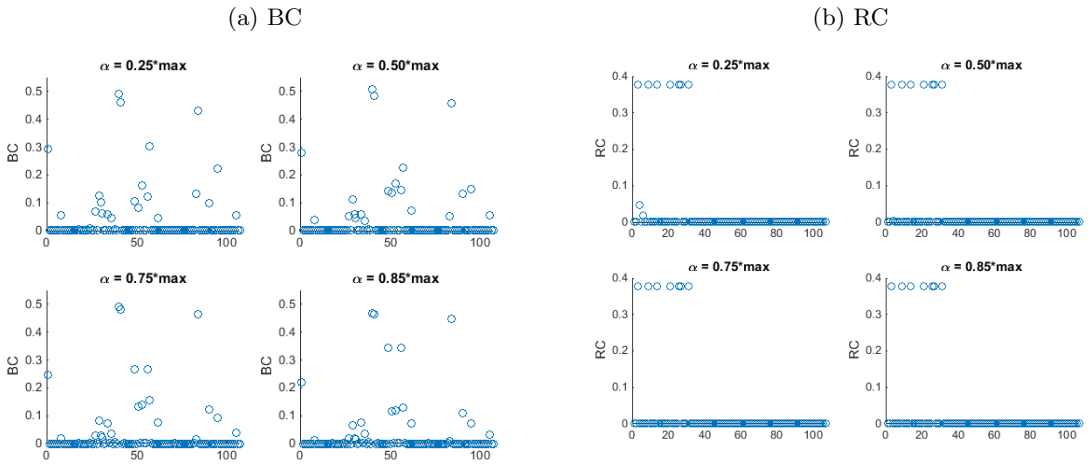


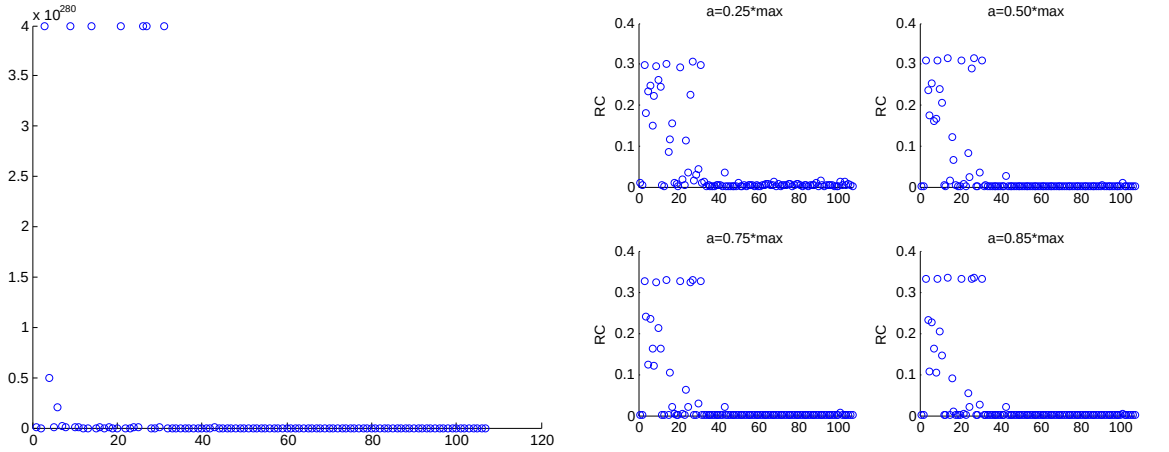
Figure 4.2a plots the RC measures for $\alpha_1 = 0.25 * \alpha_{\max}$ when Q is not normalized. (There is overflow for $\alpha = \alpha_2, \alpha_3, \alpha_4$.) The same behavior is observed, ruling normalization out as an explanation for the highly skewed distribution of the RC measures. We conclude that from a 'receiving' point of view, the same seven RN's are particularly distinct compared to the other nodes in the network, but are indistinguishable from each other. Further analysis on other attributes of these nodes is required to shed more light on why this is the case, which we leave for future work.

We point out that aggregating the base matrices $A^{[k]}$ into longer time frames resulted in better distinction among the top 20-30 RC nodes, as shown in Figure 4.2b. Note, however, that the top seven nodes remain the same, and furthermore, high RC nodes are typically staff (nodes 1-33).

Figure 4.2: RC measures (Shift 1) under different constraints. There is one data point per node; the horizontal axis is the node ID label. Nodes labeled 1-33 are staff and nodes labeled 34-107 are patients

(a) RC measures without normalization

(b) RC measures when adjacency matrices are aggregated into 20-min time frames



4.1.3 Comparison of node rankings

We consider the node rankings obtained based on BC and RC measures. Nodes are ranked from highest to lowest in descending order of the measures; the node with largest measure has rank 1. For each value of α_i for $i = 1, \ldots, 4$, we have a corresponding list of rankings l_{α_i} where

$$l_{\alpha_i}(k) := \text{ranking of node } k \text{ when } \alpha = \alpha_i.$$

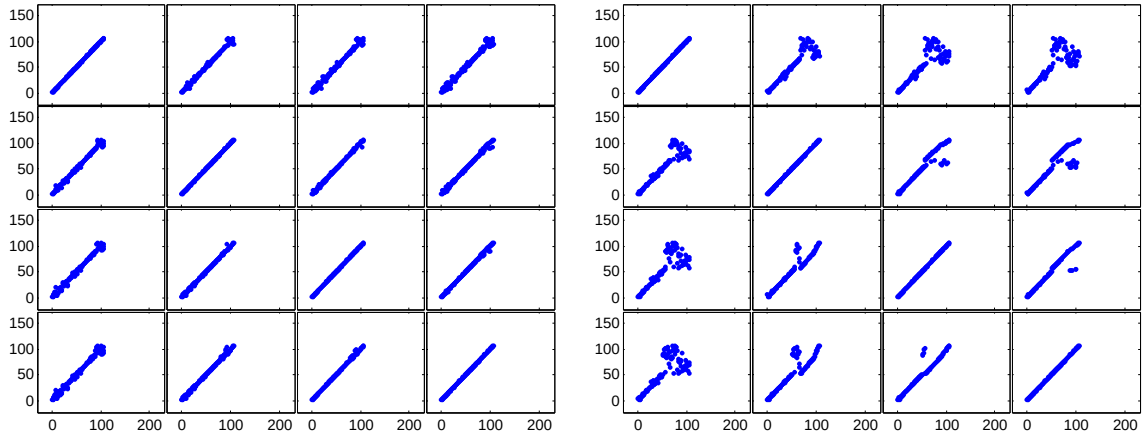
For visual reference, partial lists are shown in Appendix A. The (i, j)-th position in Figure 4.3a plots l_{α_i} versus l_{α_j}, where the rankings are based on BC measures. In Figure 4.3b we plot lists of rankings based on RC measures. The Pearson correlations corresponding to these plots are shown in Table 4.3c. We point out that Pearson correlation on these lists of rankings is the same as Spearman correlation on the lists of BC/RC measures, with the caveat that here, a large measure is associated with a small rank.

In Figure 4.4, we see that the rankings obtained for $\alpha = \alpha_1, \alpha_2, \alpha_3, \alpha_4$, are relatively robust and furthermore, the chosen values for α are far enough from the limit $\alpha \to 0$, so that the rankings are significantly different from those based on AD. Noisy behavior for low-ranked RC nodes is probably due to the small values of the RC measures: small changes can lead to drastic changes in rankings among low-ranked nodes.

Figure 4.3: Comparisons of node rankings for different values of α. Rankings according to BC are shown in 4.3a and rankings according to RC are shown in 4.3b. The (i, j)-th position plots the rankings associated with α_i versus α_j, for $i, j = 1, \ldots, 4$. Associated Pearson correlation coefficients are shown in 4.3c.

(a) BC

(b) RC



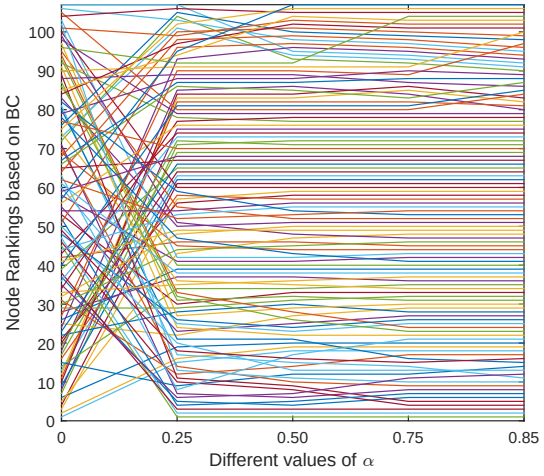
(c) Associated Pearson correlation

	α_1 v α_2	α_1 v α_3	α_1 v α_4	α_2 v α_3	α_2 v α_4	α_3 v α_4
BC	0.9946	0.9913	0.9892	0.9982	0.9968	0.9986
RC	0.9164	0.8365	0.7765	0.9422	0.8919	0.9552

From Figure 4.5 we also see that patients tend to be, on average, slightly better broadcasters than staff. As mentioned, the fact that high receivers are predominantly staff is not surprising, given their roles in the ED: Staff are well-placed to be at the receiving ends of dynamic walks.

Figure 4.4: Spaghetti plots display a line for each node connecting the rankings obtained for the different values of α. The closer the line is to horizontal, the more similar the rankings are to each other. Rankings based on aggregate degree (AD) are labeled $\alpha = 0$.

(a) BC



(b) RC

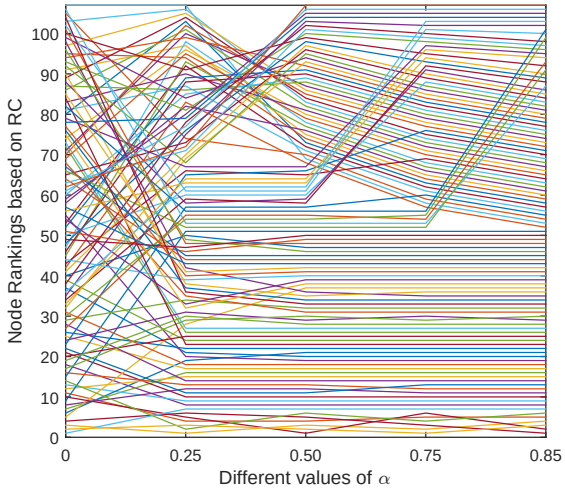
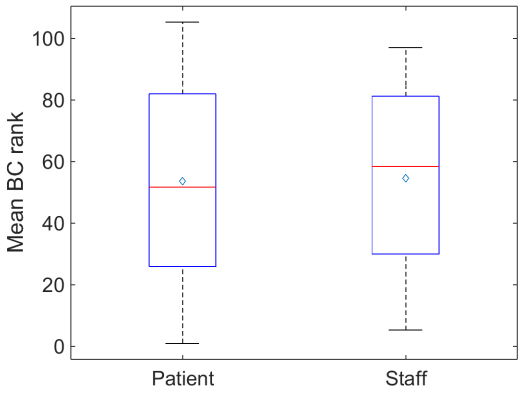
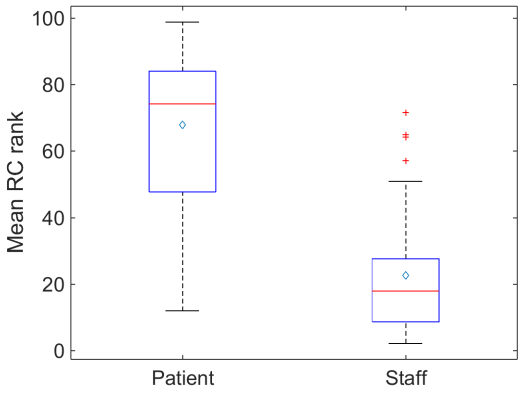


Figure 4.5: Comparison of rankings between staff and patients. We report the average ranks over $\alpha_1, \ldots, \alpha_4$. On each box, the horizontal line is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

(a) Mean BC ranks



(b) Mean RC ranks



4.1.4 Comparison of nodes in ranked order

Consider lists of nodes in ranked order. Explicitly, for each value of α_i for $i = 1, \ldots, 4$, we have a corresponding list of nodes t_{α_i} where

$$t_{\alpha_i}(k) := \text{node that has rank } k \text{ when } \alpha = \alpha_i.$$

Partial lists are shown in Appendix B. We compute Kendall correlation and intersection distance *isim* [29] as quantitative ways to assess similarity between the lists. Small values of $isim \in [0, 1]$ are indicative of strong similarity between lists. These are shown in Table 4.3 and Table 4.4.

Table 4.3: Kendall correlation between lists of nodes in ranked order.

	a_1 v a_2	a_1 v a_3	a_1 v a_4	a_2 v a_3	a_2 v a_4	a_3 v a_4
BC	0.2449	0.1963	0.1825	0.5204	0.4579	0.7366
RC	0.3871	0.2918	0.2721	0.7411	0.6805	0.8173

Table 4.4: Intersection distance (*isim*) between lists of nodes in ranked order. Values of *isim* close to 0 are indicative of strong similarity between lists.

	α_1 v α_2	α_1 v α_3	α_1 v α_4	α_2 v α_3	α_2 v α_4	α_3 v α_4
BC	0.0330	0.0518	0.0561	0.0260	0.0305	0.0060
RC	0.0749	0.0999	0.1286	0.0610	0.0784	0.0525

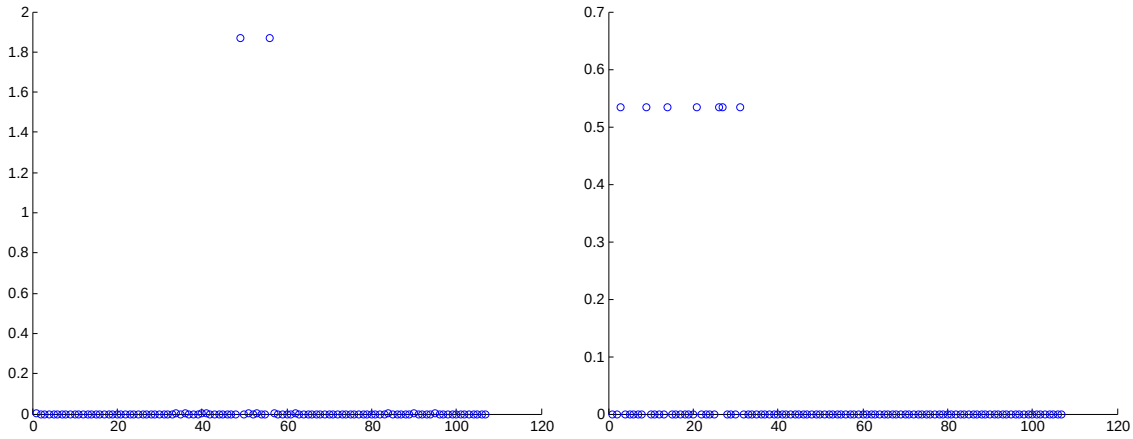
4.2 Dynamic communicability based on the matrix exponential

We compute a version of dynamic communicability based on the matrix exponential as defined in Eq. (3.4). BC and RC measures obtained are shown in Figure 4.6. We see that BC measures based on the matrix exponential are less able to distinguish between the top 20

nodes. In Figure 4.7 the resulting rankings are shown in comparison to the resolvent-based formulation of Q. The rankings obtained appear to be fairly similar, differing mostly in the low-ranking nodes.

Figure 4.6: Dynamic communicability based on the matrix exponential (Shift 1). There is one data point per node; the horizontal axis is the node ID label. Nodes labeled 1-33 are staff and nodes labeled 34-107 are staff.

(a) BC (b) RC



4.3 Convergence to aggregate degree (AD)

Recall that as $\alpha \to 0$, both BC and RC rankings should converge to AD rankings (see Section 3.2.1). We present the results for small values of α approaching zero, and in Table 4.5 we see that according to various measures, BC and RC rankings do indeed approach AD rankings[1]. This provides added assurance that in spite of the fact that the computed BC and RC measures are numerically tiny, the rankings obtained are nonetheless correct.

[1]In Table 4.5 we observe an initial drop in Kendall correlation between nodes ranked according to AD versus RC (see % $\alpha_{\max} = 0.005$). Since the correlation values are at the low end (0.25, 0.20), this anomaly is probably due to a small change in the number of discordant versus concordant pairs, and is in itself not a significant departure from the overall trend of convergence.

Figure 4.7: Comparison of node rankings between dynamic communicability based on the resolvent versus the matrix exponential

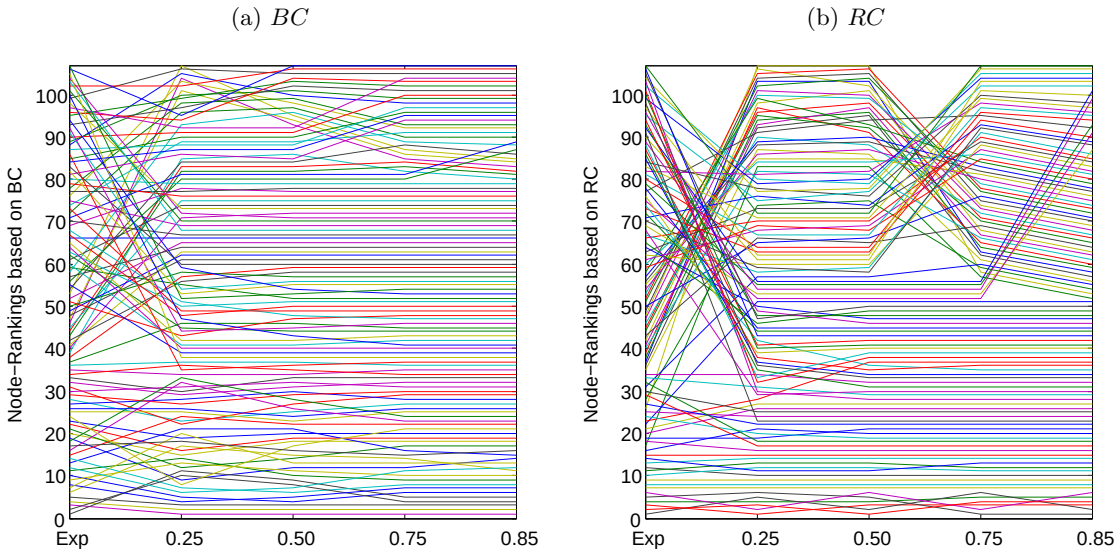
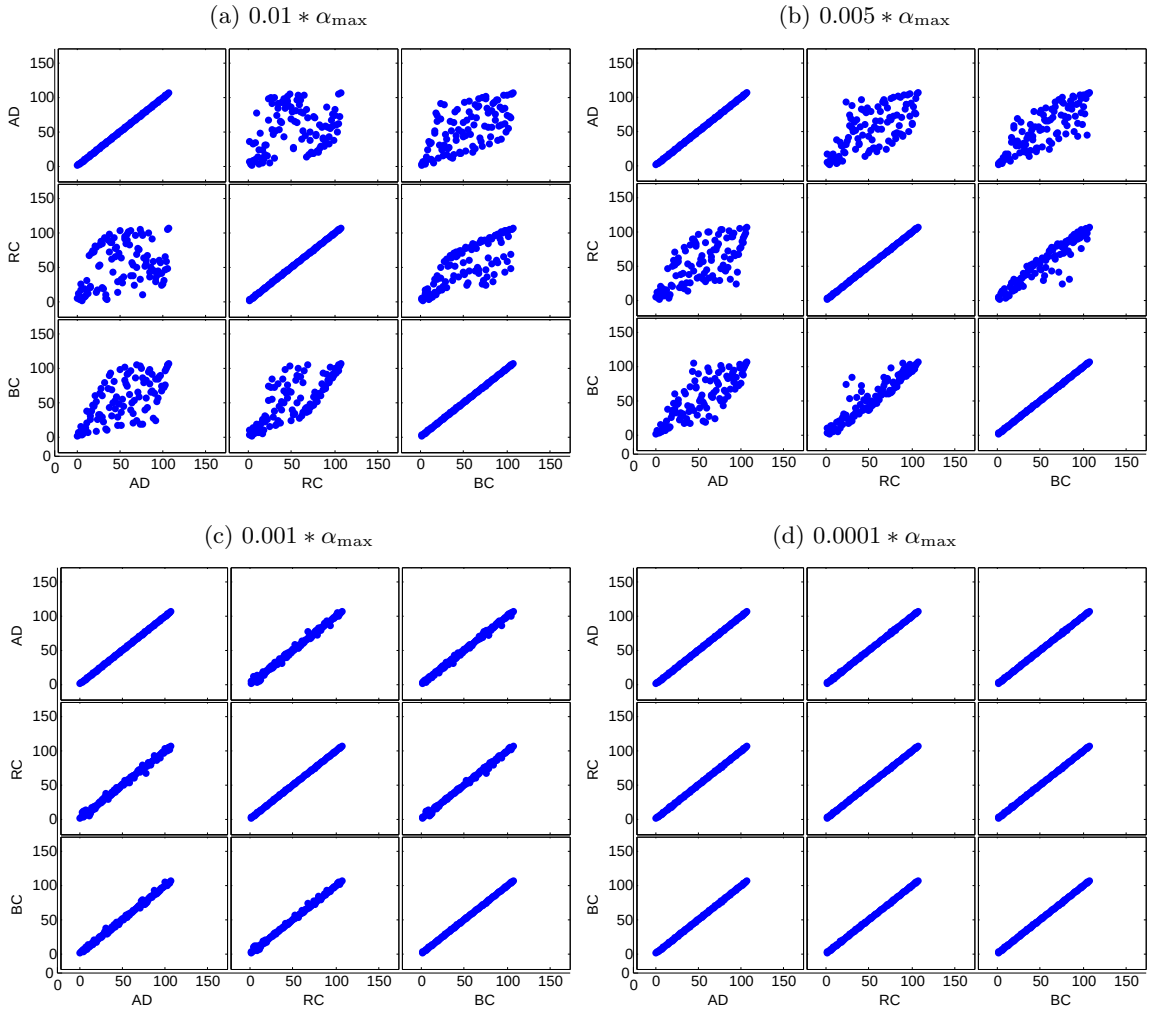


Table 4.5: Convergence to AD: Comparisons of AD, BC, RC node rankings for α approaching zero. Pearson correlation (*corr*) compares lists of node rankings. Intersection distance (*isim*) and Kendall correlation (*kcorr*) compare lists of nodes in ranked order. We also report the number of nodes in common among the top 10 and top 5.

% $\alpha_{\max}$	0.01	0.005	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
corr(AD,RC)	0.74735	0.932258	0.998119	0.999951	0.999990	0.999990	0.999990
corr(AD,BC)	0.636092	0.80269	0.997962	0.999951	0.999990	0.999990	0.999990
corr(RC,BC)	0.270381	0.675454	0.99615	0.999892	0.999980	0.999980	0.999980
Top 10 intersection	4	4	7	10	10	10	10
Top 5 intersection	2	3	3	5	5	5	5
isim(AD,RC)	0.219576	0.135918	0.034806	0.002072	0.000091	0.000091	0.000091
isim(AD,BC)	0.250046	0.178938	0.028027	0.001529	0.000123	0.000123	0.000123
isim(RC,BC)	0.364357	0.250965	0.051588	0.003602	0.000214	0.000214	0.000214
kcorr(AD,RC)	0.251984	0.19873	0.435373	0.877623	0.965086	0.965086	0.965086
kcorr(AD,BC)	0.115147	0.343326	0.348968	0.986598	0.989773	0.989773	0.989773
kcorr(RC,BC)	0.119732	0.253747	0.299242	0.864927	0.955563	0.955563	0.955563

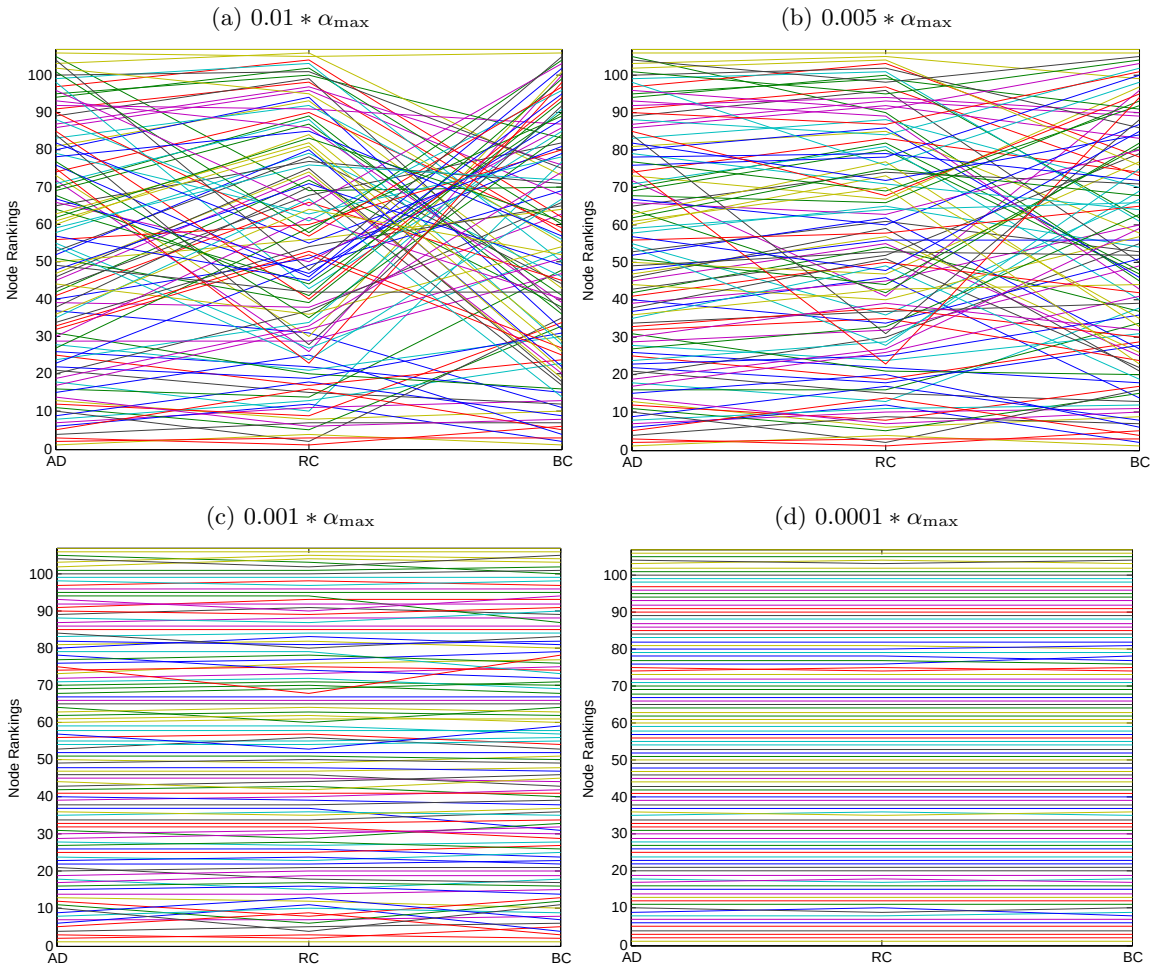
Figure 4.8: Convergence to AD: Comparisons of AD, BC and RC node rankings for α approaching zero.



4.4 Temporal dynamics of node rankings

In Figure 4.10 we show the temporal evolution of centrality-based rankings. We look at the top 10 nodes determined (at the end of the shift) by broadcast centrality (BC), receive centrality (RC), aggregate degree (AD) and binarized degree (BD). To compute the BC/RC rankings at each time step, we use Method II as discussed in Section 4.1.1. For

Figure 4.9: Convergence to AD: Spaghetti plots of node rankings based on AD, BC and RC for different values of α approaching zero. The closer the lines are to horizontal, the more similar the AD, BC and RC-based rankings are to each other.

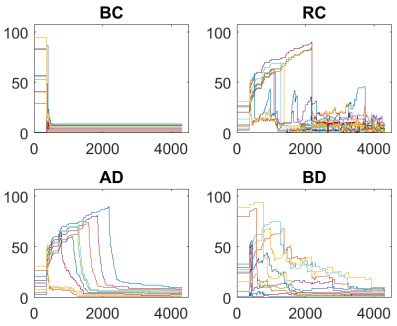


comparison, we show the results over all 7 shifts. While there are shift-specific variations, some stabilization patterns are consistently observed over all 7 shifts. BC rankings stabilize very early on in the shift, suggesting that early arrivals have a substantial advantage in developing time-respecting walks. This empirical result is in agreement with the work in [21] which showed analytically that broadcast scores increases exponentially with time. RC ranks, on the other hand, stabilize late in the shift, which means that a late arrival can still

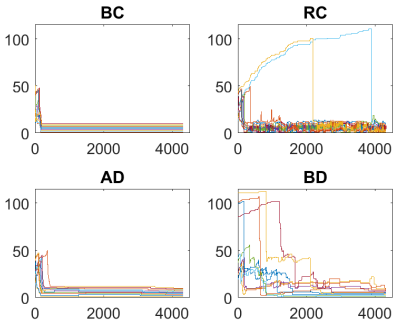
become a high receiver in a short amount of time. This duality in behavior of BC and RC rankings is also consistent with the fact that they are related by a reversal in time-ordering, as described by Eq. (3.2). AD rankings stabilize in a monotone fashion, while BD rankings are more erratic.

Figure 4.10: Dynamic rank changes of top 10 centrality nodes. Per shift we plot the temporal evolution of rankings based on broadcast centrality (BC), receive centrality (RC), aggregate degree (AD) and binarized degree (BD). Only the top 10 nodes (determined at the end of the shift) are shown.

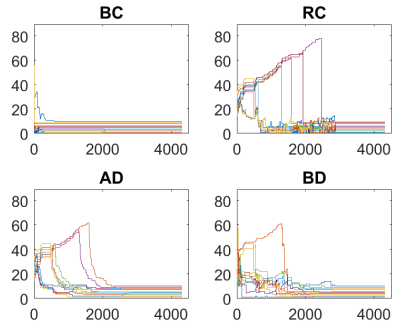
(a) Shift 1



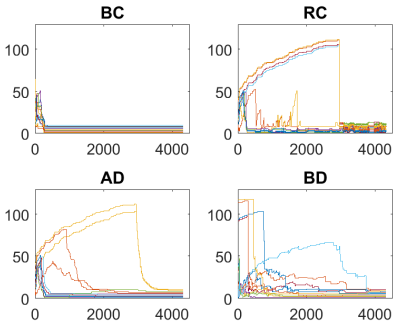
(b) Shift 2



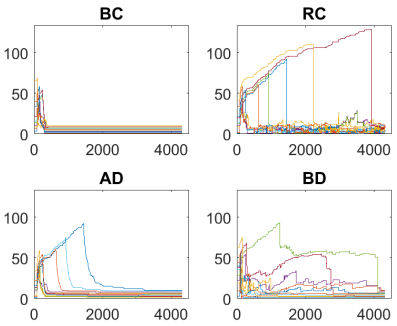
(c) Shift 3



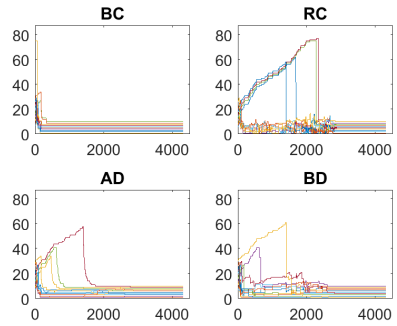
(d) Shift 4



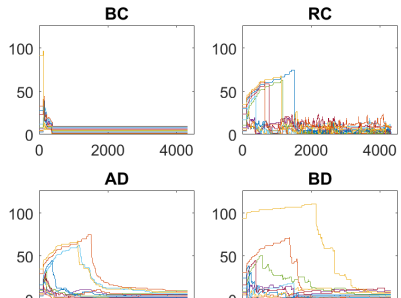
(e) Shift 5



(f) Shift 6



(g) Shift 7



Chapter 5

Interactions between top-ranked nodes

In this chapter we examine the interactions between top-ranked nodes determined by broadcast and receive centrality measures computed at the end of the shift. When and where do these top-ranked nodes meet? We work with contact data based on Shift 7 (see Table 2.1). A total of 126 participants agreed to take part in the study, out of which 35 were staff. An additional 42 patients were in the ED during that time but did not participate in the study.

Let top10BC denote the set of top 10 nodes ranked according to broadcast centrality (BC), and analogously for top10RC. Consider

$$X = \text{top10BC} \setminus \text{top10RC}$$

$$Y = \text{top10RC} \setminus \text{top10BC}$$

$$Z = \text{top10BC} \cap \text{top10RC}$$

For Shift 7, we have $X = \{3, 12, 23, 28, 33, 92\}$, $Y = \{11, 17, 24, 30, 32, 35\}$ and $Z = \{4, 5, 7, 8\}$ (not in order of centrality). There are 16 distinct nodes in $X \cup Y \cup Z$, out

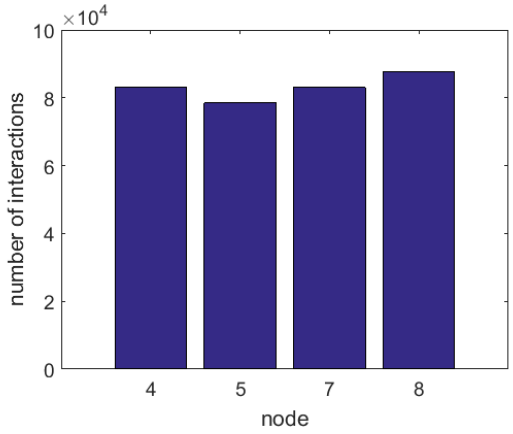
of $n = 126$ nodes in total. Note that there is only 1 patient present among these top-ranking nodes: node 92. It is also worth pointing out that nodes in Z (high broadcasters and high receivers) are staff members not including RN's (for example, adminstrative staff), while all nodes in Y apart from node 11 are RN's.

5.1 Z analysis

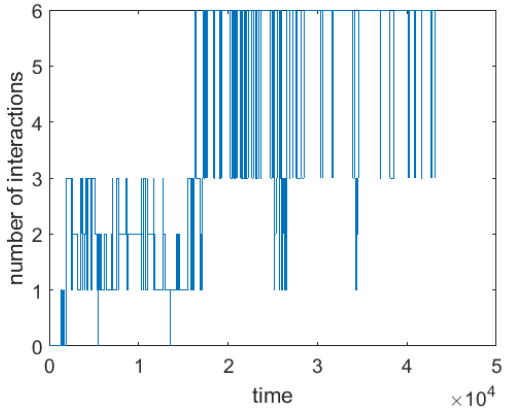
In this section we examine the interactions between nodes in $Z = \{4, 5, 7, 8\}$. Consider the network induced by Z nodes – this is the (sub)-network of interactions between Z nodes only. In Figure 5.1a we plot the degree distribution of this induced network, and we see that among this group of nodes, their interactions with each other are of a similar magnitude. Figure 5.1b plots the time-evolution of these interactions.

Figure 5.1: Interactions between Z nodes

(a) Degree of Z nodes in the induced network



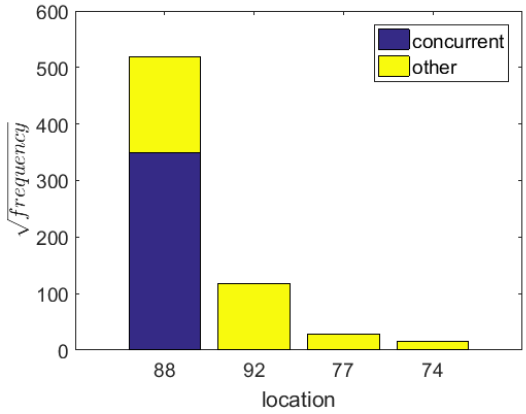
(b) Temporal evolution of induced degree



Observe that $\binom{4}{2} = 6$ is the maximum number of pairwise interactions between 4 nodes, therefore, if at any time step, if this maximum is attained, all 4 nodes must be in the same location at the same time. Such interactions are termed *concurrent*. Figure 5.1b shows that many of the Z interactions are concurrent, that is, all four Z nodes are simultaneously in

Figure 5.2: Locations of Z interactions

(a) Number of interactions per location. Concurrent interactions are those which involve all nodes in Z, and take place at the same time and place.



(b) Location description

Location	Description	Type	Area (sqft)
88	Triage & Registration	Patient Care	282
92	ED Waiting Area	Primary Waiting Area	1888
77	Staff Break Area	Administrative Support	695
74	Office Area	Administrative Support	938

the same location for much of the shift. In Figure 5.2a, we see that concurrent interactions only take place at location 88 (Triage/Registration). All other locations at which Z nodes interact are described in Table 5.2b.

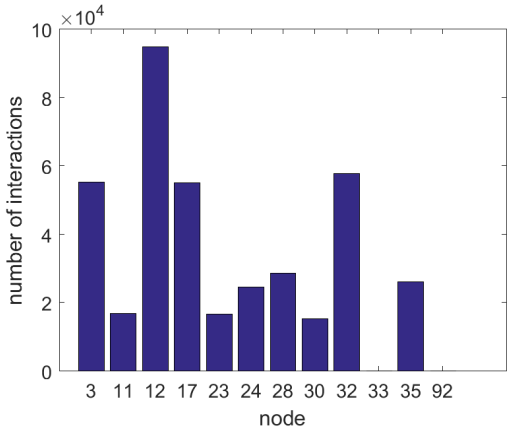
5.2 XY analysis

Recall that $X = \{3, 12, 23, 28, 33, 92\}$ and $Y = \{11, 17, 24, 30, 32, 35\}$. Consider interactions between X and Y (also denoted as 'XY interactions'). Note that we consider only edges of the form (x, y) where $x \in X$ and $y \in Y$. The *inter-group degree* of a node in X is the

number of interactions it has with nodes in Y, and similarly for the inter-group degree of nodes in Y. Figure 5.3a plots the inter-group degree of each node. Observe that nodes 33 and 92 (both in X) do not interact with nodes in Y. In Figure 5.3b we show the evolution of inter-group degree over time.

Figure 5.3: Interactions between X and Y. Only edges of the form (x, y) where $x \in X$ and $y \in Y$ are considered.

(a) Inter-group degree



(b) Temporal evolution of inter-group degree

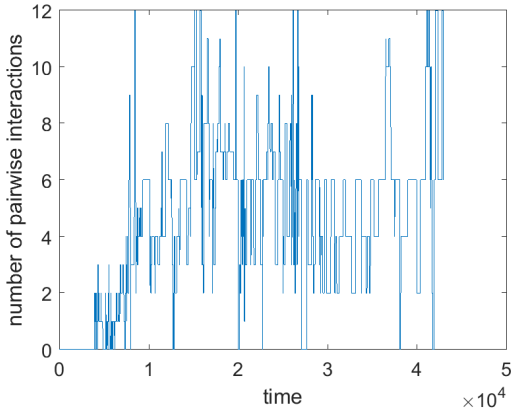
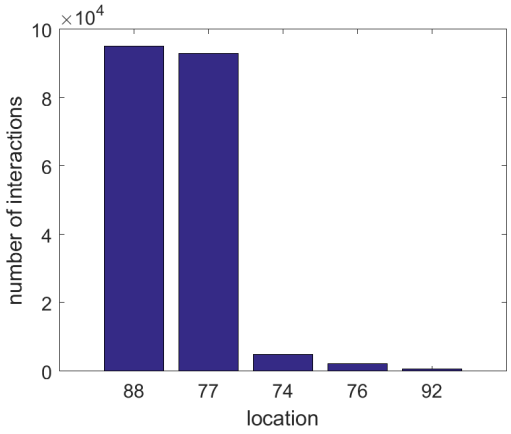


Figure 5.4a shows that the majority of XY interactions occur in location 88 (triage & registration) and location 77 (staff break area). A small number of them occur in location 92 (ED waiting area). In Figure 5.4b we plot the number of distinct nodes appearing at each of these locations. We see that while 6 distinct nodes make an appearance at location 92, they do so rarely in comparison to the other locations. Descriptions of these locations are shown in Figure 5.4c. The temporal dynamics of interactions can also be decomposed by location, as shown in Figure 5.5.

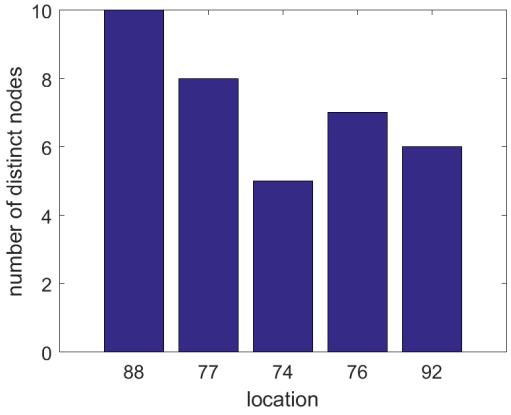
Figure 5.6a plots the frequency of the number of interactions at each time step. We see that at each time step, there are at most 12 interactions, while groups of 6 interactions occur the most frequently. The interactions shown in Figure 5.6a take place at the same time step, but may be spread out over different locations. For example, suppose there are

Figure 5.4: Locations of XY interactions

(a) Number of interactions per location



(b) Number of nodes per location



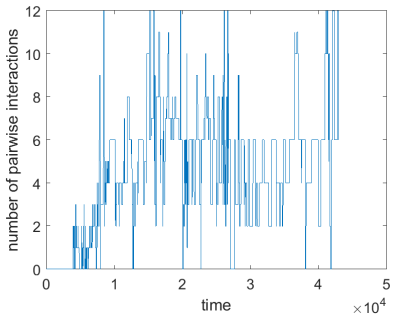
(c) Location description

Location	Description	Type	Area (sqft)
88	Triage & Registration	Patient Care	282
77	Staff Break Area	Administrative Support	695
74	Office Area	Administrative Support	938
76	Office Area	Administrative Support	708
92	ED Waiting Area	Primary Waiting Area	1888

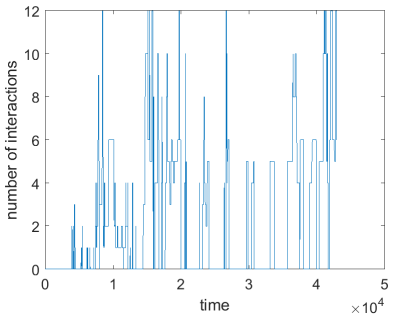
6 interactions at time t: 4 of these may be in location 88 while 2 are in location 74. To get a better idea of how these interactions are spread out in the ED, we consider groups of interactions of the same count (from 1 to 12), which in addition, **a)** take place over a contiguous length of time I; **b)** involve the same group of nodes throughout I; and **c)** occur in the same location (which may change during I). Such groups of interactions can be viewed as 'communities', since they involve a fixed group of nodes and take place in the same location, over an extended length of time I. Note that we do not take into consideration the length of interval I: every contiguous time interval, during which a fixed group of nodes interacted, is treated the same, regardless of length. The distribution of

Figure 5.5: Temporal dynamics of XY interactions per location

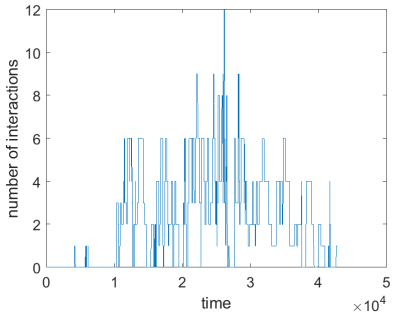
(a) All locations



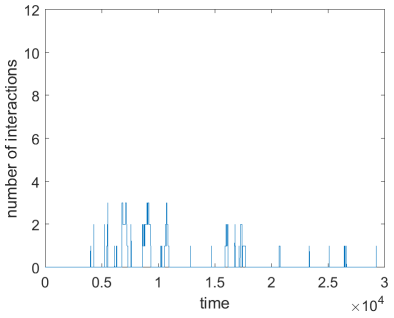
(b) 88: Triage & registration



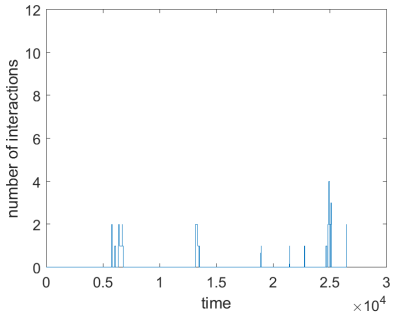
(c) 77: Staff break area



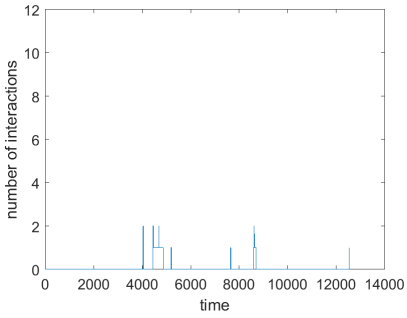
(d) 74: Office area



(e) 76: Office area



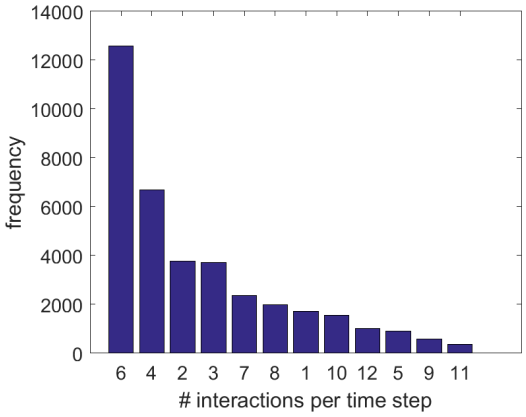
(f) 92: ED waiting area



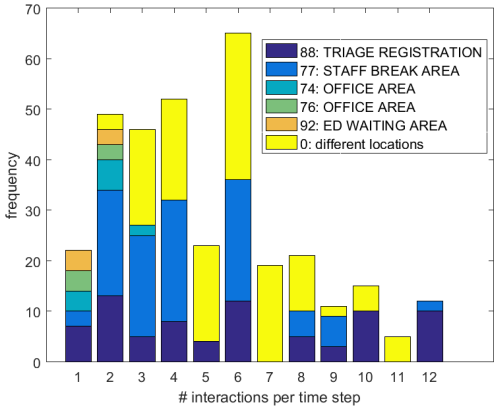
such communities per location is shown in Figure 5.6b. This shows us that, for example, out of all 6-interaction groups, almost half take place in different parts of the ED.

Figure 5.6: XY-group interactions over time and space

(a) Distribution of interactions per time-step



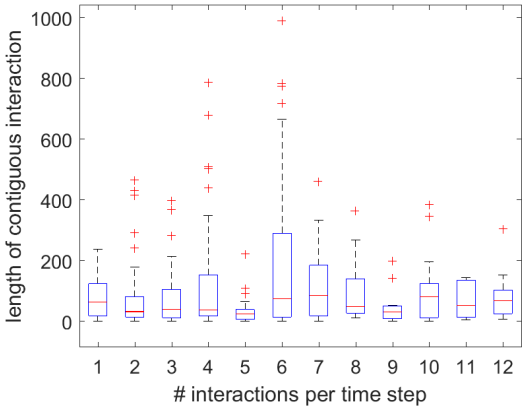
(b) Distribution of 'communities' per location



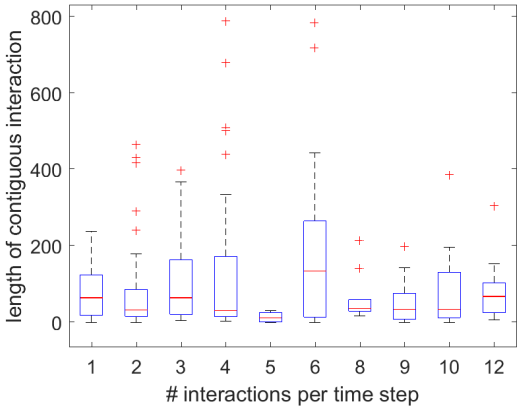
It may also be of interest to examine the lengths of contiguous time-interval I associated with each interaction-group, as shown in Figure 5.7. In Figure 5.7a, the locations associated with these interactions are not unique, while in Figure 5.7b, we consider only group-interactions that take place at a unique location. For example, Figure 5.7b shows us that among 'communities' involving 6 interactions, the median length of contiguous interaction time is about 180 time-steps.

Figure 5.7: Distributions of contiguous interaction time. On each box, the horizontal line is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

(a) Group-interactions irrespective of location



(b) Location-specific group-interactions



Chapter 6

Measures of virulence

We simulate contagion processes on the contact network described in Chapter 2 and examine the epidemic effect associated with each node as the initial source of infection. Our approach therefore departs from typical contagion modeling on networks, which aims to quantify the spreading capability of the network as a whole. (Often, a subset of nodes is chosen at random from which the infection is seeded [65, 31].) More importantly, the dynamic nature of the contact information as described in Chapter 2 allows us to explicitly utilize empirically observed interactions in the contagion process. This is in contrast to traditional epidemic processes on static networks [32, 57, 61], where edges representing potential disease-spreading contacts are considered present and unchanging over time. Consequently, assuming there is no recovery, it is only a matter of time before everyone in (connected components of) the network is infected. However, in reality, contacts themselves form and dissolve over time, and given this information, we hope to paint a more realistic picture of how contagion spreads on a network [8, 81].

Epidemic modeling on networks typically also involves updating the infection status of nodes after each time step ([31, 82, 32, 57, 61]). Each time step is treated independently from all other time steps, and the probability of infection is assumed to be constant and

unchanging over time. Consequently, all contact history between nodes is disregarded in the infection process, leading to an inability to distinguish pairs of nodes which have just come into contact, from pairs of nodes which have already been in contact for a long period of time. We argue that whether or not a pair of nodes has been in contact in the previous time-step should make a difference to the current infection probability, since it is reasonable to assume that the longer two nodes are in contact with each other, the more likely they are to spread disease, share ideas, or rumors, or influence each other in some way. We therefore seek to model contagion by explicitly incorporating the length of contact time between nodes.

We consider various infection processes on the network. First, we model contagion as a stochastic process, where the probability of infection depends explicitly on the length of contact time between nodes. This is explained in detail in Section 6.1. In this approach, infection essentially spreads at a non-uniform rate: infection can sometimes spread after a short amount of contact time, albeit with low probability. This resulted in a wide range of observed epidemic outcomes, with no characteristic shape associated with the distributions (per initial source). We therefore consider also a deterministic approach, where infection spreads after a fixed, pre-determined amount of uninterrupted contact time, which we refer to as the *infection rate*. Specifically, we study the epidemic effect when contagion spreads after 10, 15, 20, 25 minutes of contiguous contact time. This allows us to get a better sense of how epidemic outcome depends on contact time between nodes. We seek to answer the question: what is the epidemic effect when infection occurs after a fixed amount of contact time? This is in contrast to the stochastic process, which in essence provides an overview of what happens over all possible infection rates. For comparison, we also consider the worst-case scenario, where every 10-second contact spreads the infection. The results associated with the different infection processes are shown in Section 6.2.

We emphasize that the contagion process on the network is completely independent of

the notion of dynamic communicability as discussed in Chapter 3. While both depend explicitly on the edge structure of the network, the parameters involving the infection process are independent from the parameters involving dynamic communicability. The aim of this work is to examine the relationship between these two independently-defined notions of dynamic communicability and contagion.

6.1 Stochastic Model

In this section we describe the stochastic approach used to model the spread of disease on the dynamic network, where the length of contact time between nodes is explicitly used to compute the probability of infection. Parameters for infection are chosen based on rates observed for influenza [55, 62, 69]. Within the framework of a limited observation period, we assume that infected nodes are immediately infectious. In addition, since recovery is not physically feasible within this time frame, the Susceptible-Infected (SI) model is adequate, and models the early phase of an outbreak. For each simulation, there is only one initial source of infection, which is infectious upon its arrival in the ED.

We assume that infection between susceptible-infected pairs is a Poisson process, which is, in particular, independent and memoryless. Consequently, the time to infection, X, follows an exponential distribution. We write $X \sim \text{Exp}(\lambda^*)$. The parameter λ^* is chosen to satisfy

$$\Pr(X \leq 60 \text{ sec} \mid \lambda^*) = \int_0^{60} \lambda^* e^{-\lambda^* t}\, dt = 1 - e^{-60\lambda^*} = 0.009,$$

where the value of 0.009 is chosen based on an approximated attack rate observed in an outbreak of influenza aboard a commercial airliner [55, 62, 69]. Solving for λ^*, we obtain $\lambda^* \approx 1.5 \times 10^{-4}$. Note that $1/\lambda^* \approx 664$ seconds is interpreted as the average time to infection. Suppose nodes i and j are in contact for t_{ij} seconds, then the probability of

disease transmission is given by

$$p_{ij} = \int_0^{t_{ij}} \lambda^* e^{-\lambda^* t}\, dt.$$

In order to determine if infection takes place, a random number $u \in \text{Unif}(0, 1)$ is generated. If $u < p_{ij}$, infection occurs, otherwise, infection does not occur. The rationale here is that for 'large' p_{ij}, we would like infection to occur as much as possible. Note that p_{ij} is computed based on the entire length of uninterrupted contact time between nodes i and j. Infection, and therefore further spreading potential, occurs not at the end of the contact period t_{ij}, but at the time-step when p_{ij} exceeds u. Note that if the pair (i, j) comes into contact multiple times over the course of study, each contiguous contact period is treated independently.

We gratefully acknowledge the contribution of Andres Celis who implemented the infection process in Java. Pseudo-code can be found in Appendix C. The output of the Java code is stored in csv format, which is then analyzed in Matlab.

6.2 Results

We present the results based on the temporal network of Shift 7 ($n = 126$). Each node is considered in turn as the initial source of infection; per initial source, we look at the total number of infections, or final epidemic size (EPI), that occur by the end of the shift.

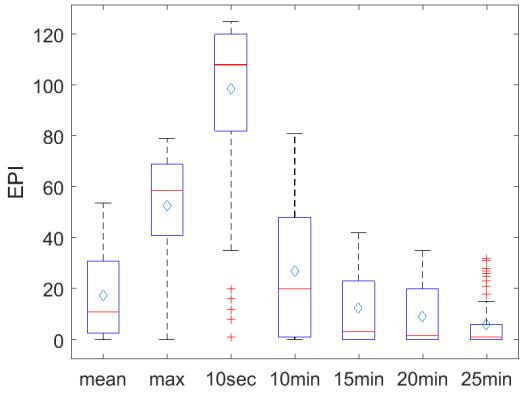
For the stochastic process, we repeat the simulations $N = 1000$ times per initial source. Observed distributions of EPI per seed node have no characteristic shape and are typically not symmetric nor unimodal. We consider both the mean and maximum values as summary measures of virulence. Note that the maximum epidemic size is conditional on N, but since all nodes are subject to the same number of simulations, we can ignore the conditional in the following discussion.

In contrast to the stochastic process, suppose that contagion spreads only after uninterrupted contact over T minutes, where $T = 10, 15, 20$ and 25. We also consider the worst-case scenario, where every 10-second contact spreads the infection. These are collectively referred to as different 'infection strategies'.

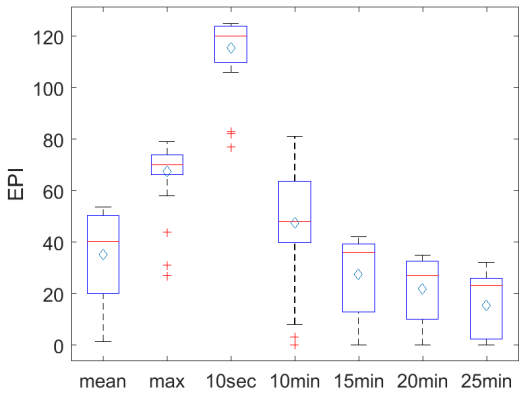
Figure 6.1 displays boxplots associated with all infection strategies. In Figure 6.1b we show the distribution of EPI associated with staff members as initial source, while Figure 6.1c shows the distribution of EPI associated with patients as initial source. We see clearly that staff members as infection seeds are associated with larger epidemic outcome, regardless of infection strategy. Unsurprisingly, in the worst-case scenario, where every 10-second contact spreads the disease, a large proportion of the population becomes infected by the end of the shift. The stochastic worst-case scenario (labeled 'max') was less severe in comparison, but still infected many more than if every uninterrupted 10-minute contact was contagious. Overall, in Figure 6.1a we see that stochastic worst-case epidemics do not affect more than 60% of the population under study, while on average (stochastically), less than 24% of the population become infected.

Figure 6.1: Comparison of epidemic outcomes between staff and patients. On each box, the horizontal line is the median, the diamond indicates the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. In all cases, staff are associated with larger epidemic outcome than patients.

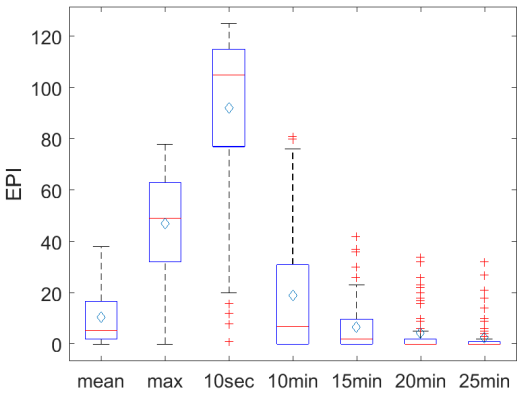
(a) All



(b) Staff



(c) Patients



Chapter 7

Relationship between centrality and virulence

The over-arching goal of this work is to examine the relationship between the network-based centrality score and epidemic outcome associated with the node which seeded the infection. We are particularly interested in temporal centrality as captured by dynamic communicability: how does the broadcasting ability of a node, quantified by broadcast centrality (BC), relate to epidemic outcome? For comparison, we also study the following centrality measures in decreasing temporal complexity: aggregate degree (AD) and binarized degree (BD) as described in Section 3.3.

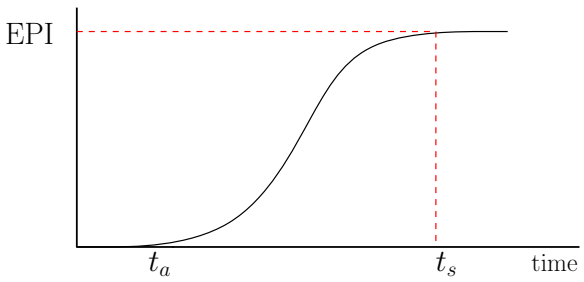
Measures of virulence under consideration are the stochastic mean and maximum epidemic size obtained from the infection simulations discussed in Section 6.1. We also consider the non-stochastic worst-case epidemic size, where every contact spreads the infection. A schematic of the number of infections (associated with a single seed node) over time is shown in Figure 7.1. Since there is no recovery, this is a non-decreasing count over time.

We associate with each seed node a (non-stochastic) epidemic measure of the form

$$\text{NS-EPI} = \frac{\text{EPI}}{\log(t_s - t_a)}$$

where EPI is the final epidemic size associated with the seed node, t_a ('activation' time) is the time at which an infection first spreads and t_s ('saturation' time) is the earliest time at which the final epidemic size is reached. The rational for the choice of measure is this: we want the measure to be proportional to the final epidemic size, but inversely proportional to the length of time it took to spread the infection. In other words, among nodes associated with the same final epidemic size, those which required less time to infect the same number of people will be associated with a larger epidemic measure.

Figure 7.1: Schematic of worst-case epidemic size associated with a single seed node



We first compare the rankings of the nodes based on centrality, with the rankings based on epidemic outcome. Throughout this work, a node with the largest centrality/epidemic measure is ranked number 1. (Note that this is in contrast to the traditional definition of Spearman rank correlation, where measures are typically ranked in increasing order, so that the smallest measure has rank 1.) Rank correlations are used to measure the strength of the *overall* relationship between the two ranking systems. We then focus attention on only the highly-ranked nodes – how well are centrality rankings able to capture highly virulent nodes, and vice versa? Next, we look at how the *measures* themselves are related. We use the tools of linear regression to quantify the tendency of epidemic outcome (the 'response' variable)

to vary with centrality (the 'predictor' of interest) in a statistic fashion. Linear regression techniques allow us to answer questions of the form: what is the effect on epidemic outcome as the centrality of a node increases? The framework also allows us to account for the effect of so-called 'confounding variables' which may obscure the true relationship between the variables of interest. In Section 7.2 we explain and motivate our choice of variables included in the statistical model, before discussing the parameter estimates obtained. Interaction effects between the other variables and centrality are presented in Section 7.3. Finally, to test the efficacy of the regression model in explaining epidemic outcome, in Section 7.4 we use the parameter estimates derived from Shift 1 data to *predict* the epidemic outcomes associated with Shift 2 to 7. Prediction errors form another means with which to assess the relevance of centrality as a predictor of epidemic outcome.

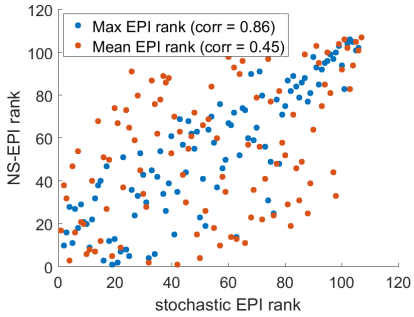
7.1 Ranks analysis

In this section we examine the relationship between node rankings based on epidemic outcome, and node rankings based on network centrality. We reiterate that a small numerical value is associated with a high rank: a node with rank 1 is the most central/virulent. Figure 7.2 plots the epidemic-rankings based on the stochastic simulations (mean/max EPI) relative to the rankings based on the non-stochastic epidemic measure NS-EPI; Figure 7.2h reports the rank correlations averaged over 7 shifts. As expected, the rankings based on NS-EPI correlate strongly with the rankings based on max EPI. This is not surprising, since both are versions of the worst-case scenario.

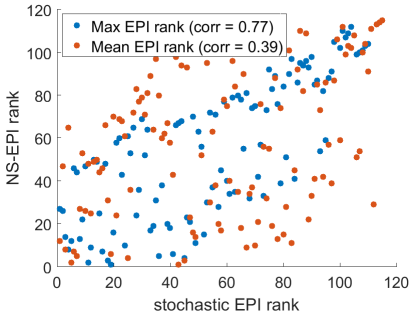
In Figure 7.3 we plot the relationship between rankings based on the stochastic mean epidemic size (mean EPI) relative to the network-based centrality rankings. On the left, we consider the temporal centrality measures BC and RC, while on the right, we look at the less nuanced measures, BD and AD. For simplicity we refer to both BD and AD as 'static' measures, even though AD contains some temporal information. The same format is used in

Figure 7.2: Comparison of EPI rankings based on stochastic and non-stochastic measures. NS-EPI is the measure defined by $EPI_w/\log(t_s - t_a)$, where EPI_w is the final epidemic size associated with the worst-case scenario (every contact spreads the disease); t_s is the earliest time at which this maximum was attained, and t_a is the time at which the first infection occurred.

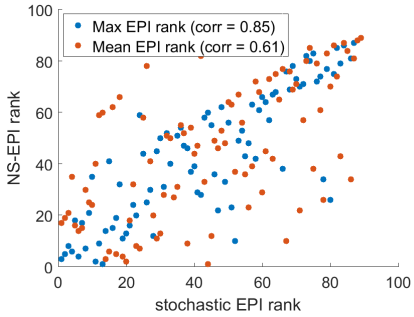
(a) Shift 1



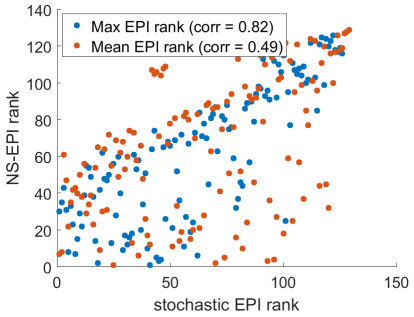
(b) Shift 2



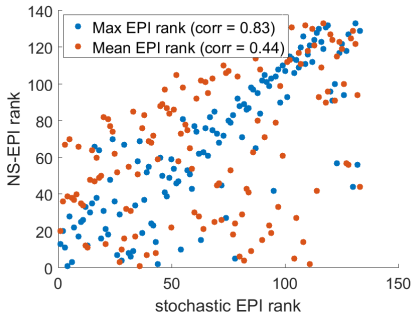
(c) Shift 3



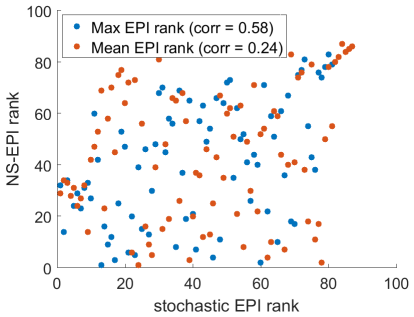
(d) Shift 4



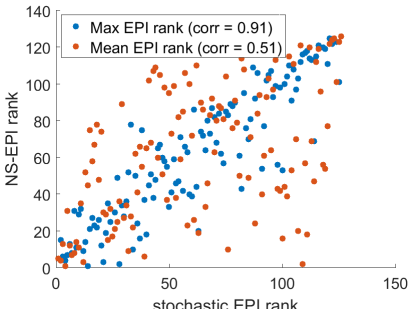
(e) Shift 5



(f) Shift 6



(g) Shift 7



(h) Correlations averaged over 7 shifts

	mean EPI	max EPI
NS-EPI	0.45	0.80

Figures 7.4 and 7.5, where we plot the rankings based on the stochastic maximum epidemic size (max EPI) and the non-stochastic epidemic measure (NS-EPI) respectively. We show results for Shifts 1 and 7 only; we refer the reader to Appendices D, E and F for the results on all other shifts. It is interesting to note that mean EPI rankings have a stronger relationship with the static measures (BD and AD), but with respect to both versions of the worst-case scenario (max EPI and NS-EPI), the temporal measure BC significantly outperforms BD and AD. In fact, for both Shifts 1 and 7, BC ranks are almost perfectly correlated with NS-EPI ranks (Figures 7.5a, 7.5c), although for Shift 1, the relationship between the top-ranking nodes is slightly weaker. In Table 7.1 we show the rank correlations averaged over 7 shifts. While RC rankings do not have a strong overall relationship with EPI rankings, from Figures 7.4 and 7.2 we see that high receivers are sometimes associated with large epidemic outcome. The same is true for BD and AD ranks: while overall correlation with EPI ranks are not strong, restricting attention to top-ranking nodes paints a different picture. These observations highlight the fact that summary measures such as correlation should be interpreted alongside the visual picture as much as possible.

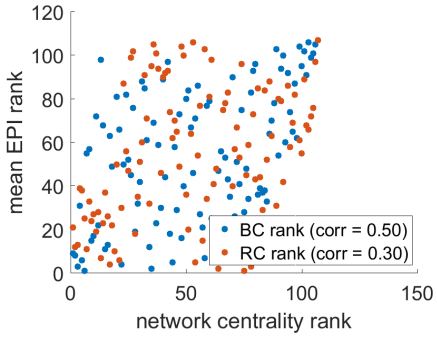
This motivates us to look at how the different centrality measures perform with respect to identifying highly virulent nodes. We look at the centrality measures in increasing temporal complexity: BD, AD and BC. Figure 7.6 shows that for Shifts 1 and 7, BC is able to identify virulent nodes that are not captured by BD and AD – these nodes are dynamic communicators in the sense that they are not identified as important by aggregated or non-temporal measures such as AD and BD respectively, but are ranked highly according to temporal measures such as BC (see Section 3.2). Utilizing all three measures together thus increases the coverage of top-spreaders. We quantify the added-value of BC in the identification of top-spreaders, by computing the percentage of dynamic communicators relative to the number of virulent nodes captured by all three measures. The added-value of BC averaged over 7 shifts is shown in Table 7.6e. All values are larger than zero, which

Table 7.1: Rank correlations averaged over 7 shifts

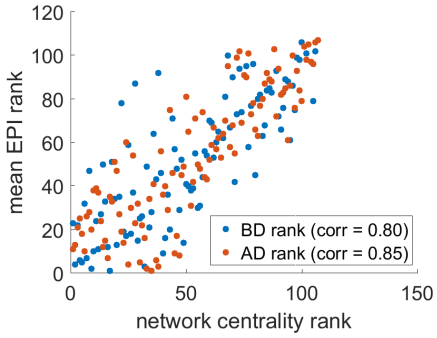
	Mean EPI	Max EPI	NS-EPI
BD	**0.86**	0.61	0.34
AD	**0.89**	0.53	0.28
BC	0.56	**0.87**	**0.94**
RC	0.36	-0.04	-0.26

Figure 7.3: Comparison of mean EPI rankings and network-based centrality rankings. Network-based centrality measures under consideration are the temporal measures, BC and RC, as well as the less nuanced measures, BD and AD. For simplicity, we refer to BD and AD as 'static' measures, even though AD contains some temporal information. Shifts 1 and 7 are shown; for the other shifts, see Appendix D.

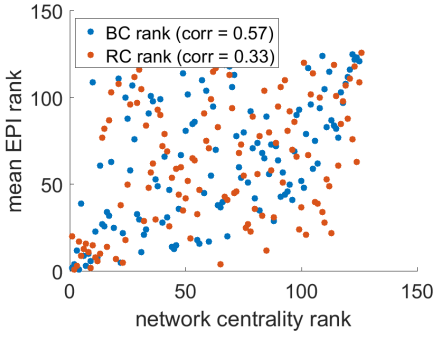
(a) Shift 1: Temporal



(b) Shift 1: Static



(c) Shift 7: Temporal



(d) Shift 7: Static

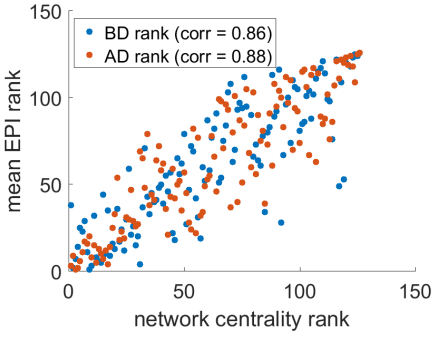
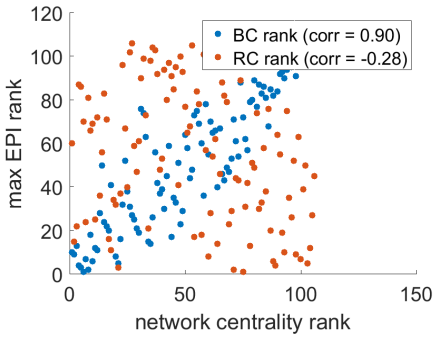
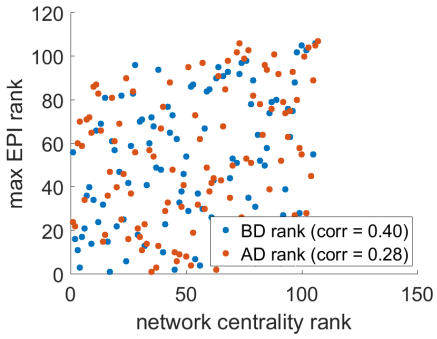


Figure 7.4: Comparison of max EPI rankings and network-based centrality rankings. Network-based centrality measures under consideration are the temporal measures, BC and RC, as well as the less nuanced measures, BD and AD. For simplicity, we refer to BD and AD as 'static' measures, even though AD does contain some temporal information. Shifts 1 and 7 are shown; for the other shifts, see Appendix E.

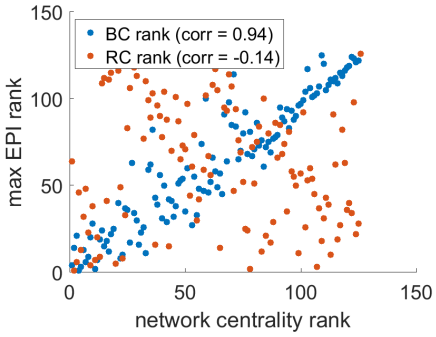
(a) Shift 1: Temporal



(b) Shift 1: Static



(c) Shift 7: Temporal



(d) Shift 7: Static

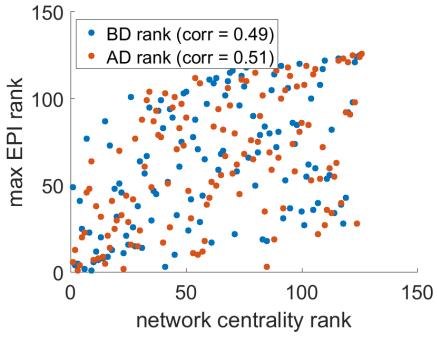
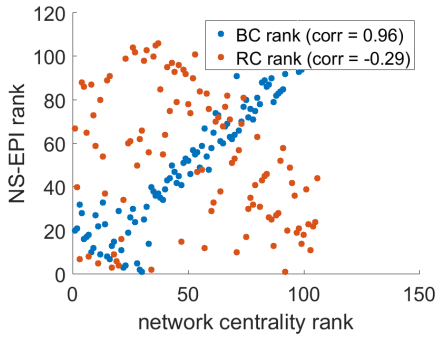
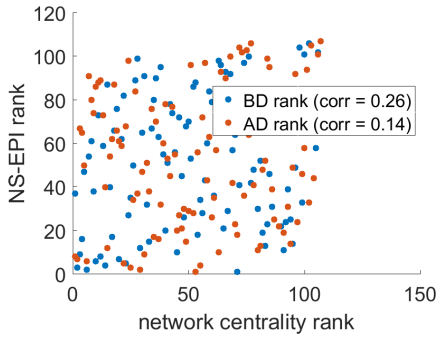


Figure 7.5: Comparison of NS-EPI rankings and network-based centrality rankings. Network-based centrality measures under consideration are the temporal measures, BC and RC, as well as the less nuanced measures, BD and AD. For simplicity, we refer to BD and AD as 'static' measures, even though AD does contain some temporal information. Shifts 1 and 7 are shown; for the other shifts, see Appendix F.

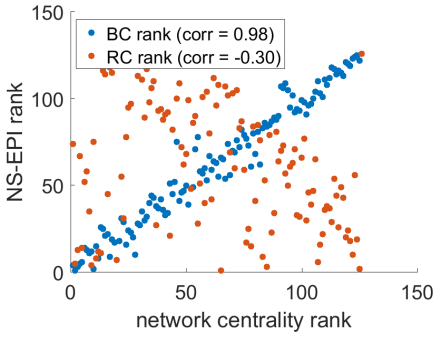
(a) Shift 1: Temporal



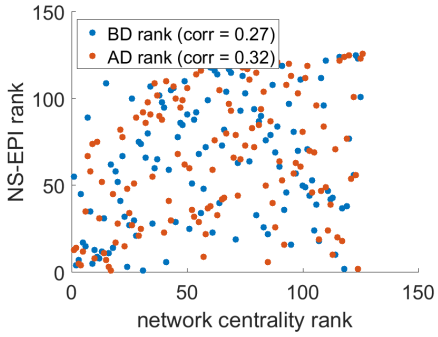
(b) Shift 1: Static



(c) Shift 7: Temporal



(d) Shift 7: Static



means that BC always adds some value in identifying virulent nodes. Again, we see that BC performs better with respect to worst-case epidemics.

To complete our analysis, we look at the average centrality rankings of virulent nodes, as well as the average EPI-rankings of central nodes, shown in Tables 7.2 and 7.3. Highlighted in bold are the smallest average rank (over the 4 centrality measures) associated with each epidemic measure. Table 7.2 shows that the nodes associated with the top 10 largest stochastic mean epidemic size have average AD rank of 12.87, and this is the smallest average over all centrality measures, suggesting that AD is the best indicator for nodes with large mean epidemic size. Among all centrality measures, high BC nodes are associated with the largest maximum epidemic size and non-stochastic epidemic measure, reiterating the observation that BC is a strong indicator of worst-case epidemics. While central nodes consistently rank highly with respect to mean epidemic size (Table 7.3 – mean EPI), this is not the case for max EPI and NS-EPI. With respect to worst-case epidemics, there is more variability in the ability of the different centrality measures to mimic the EPI-rankings. We see again that BC performs the best in this context.

7.2 Regression analysis

In this section we use linear regression techniques to quantify the overall relationship between network-based centrality measures and epidemic outcome. Analysis is performed on Shift 1 data. We reiterate that here, the *numerical measures* are used, in contrast to Section 7.1, where the analysis is performed on the *rankings* based on the centrality scores. The aim is to quantify the statistical relationship between epidemic outcome and the attributes of the initial source of the infection. The attribute of interest is its centrality score, and in particular its broadcast centrality (BC) score. For comparison, we also consider other measures such as binarized degree (BD) and aggregate degree (AD). Observe that if a node enters the ED late in the shift, or is present in the ED for a short amount of time, it will

Table 7.2: Average centrality rankings of top 10 EPI nodes

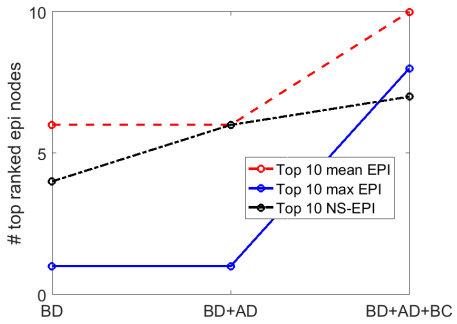
	BD	AD	BC	RC
top 10 mean	14.31	**12.87**	14.13	26.46
top 10 max	24.56	29.97	**12.03**	50.03
top 10 NS-EPI	40.97	46.73	**17.79**	59.37

Table 7.3: Average EPI rankings of top 10 centrality nodes

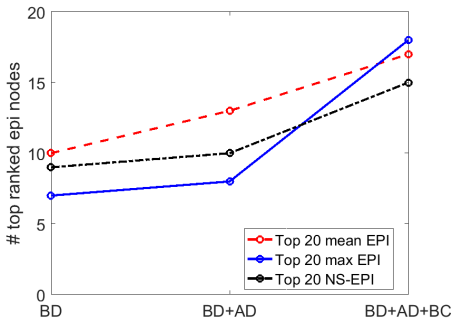
	mean EPI	max EPI	NS-EPI
top 10 BD	16.50	21.29	29.66
top 10 AD	**12.53**	28.64	41.13
top 10 BC	16.60	**12.79**	**20.91**
top 10 RC	16.64	38.97	51.29

Figure 7.6: Identification of top spreaders. The vertical axis counts the number of nodes ranked highly in terms of both epidemic outcome and centrality (horizontal axis). Including centrality measures in order of complexity increases the coverage of virulent nodes. See Appendix G for other shifts. Table 7.6e shows the added-value of BC in identifying top spreaders, where the added-value of BC is quantified by the percentage of dynamic communicators relative to the total number of virulent nodes captured by all three measures.

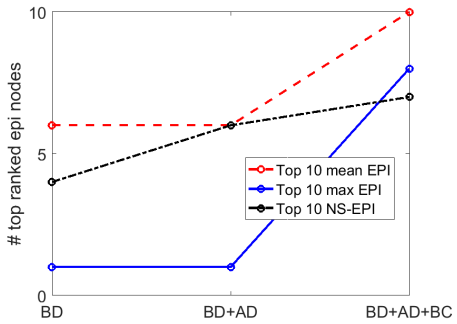
(a) Shift 1: Top 10



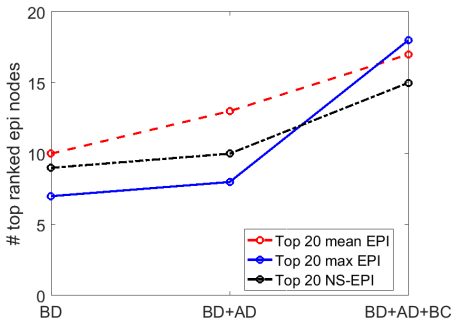
(b) Shift 1: Top 20



(c) Shift 7: Top 10



(d) Shift 7: Top 20



(e) Added-value of BC averaged over 7 shifts. A positive value means that there exist virulent nodes that are identified by BC, but not by BD or AD.

	Mean EPI	Max EPI	NS-EPI
Top 10	0.14	0.32	0.29
Top 20	0.06	0.28	0.34

have less opportunity to develop connections and consequently, walks across the network, and its walk-based centrality scores (such as BC) are likely to suffer as a result. Therefore, possible confounding factors such as the time of first appearance in the ED (T) and duration observed in the ED (D) are included in the analysis, eliminating the need to employ the use of sliding windows as is common in the literature [53, 65]. Staff/patient category is also included as a predictor in the model.

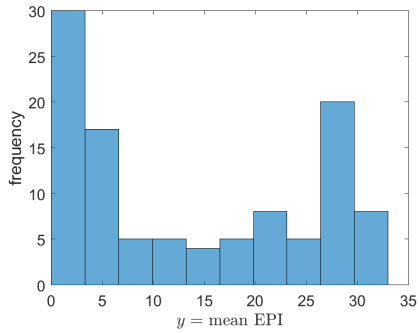
As in Section 7.1, we consider the following three epidemic outcomes associated with each initial source node: the stochastic mean epidemic size (mean EPI), the stochastic maximum epidemic size (max EPI) and the non-stochastic epidemic measure (NS-EPI). We first examine the normal assumption underlying the linear regression model. Box-Cox transformations of the form $\frac{y^\lambda - 1}{\lambda}$ can be used to transform non-normal data to data that has an approximately normal distribution. We use Matlab's `boxcox` function, which approximates the value of λ that maximizes the log-likelihood function and transforms the data accordingly. The left-hand column of Figure 7.7 shows the distribution of each of the response variables; the right-hand column shows the corresponding Box-Cox-transformed distributions. We see that the raw response data is not normally distributed, and while the Box-Cox transformation slightly evens out the distribution of max EPI, it does not change the shape of either mean EPI or NS-EPI. Consequently, for the remaining analysis we will use the raw response data.

In the following analysis we use BC measures[1] associated with $\alpha = 0.25 * \alpha_{\max}$. Figure 7.8 displays the relationship between the response variables and the predictor of interest, log(BC). We see that log(BC) correlates very strongly with both max EPI and NS-EPI. The relationship with mean EPI is less clear: there exists some nodes with relatively large

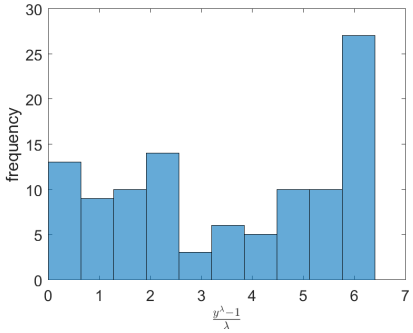
[1] As discussed in Section 4.1, the BC and RC rankings obtained are fairly robust over the range $\alpha \in [0.25, 0.85] * \alpha_{\max}$ studied. Upon closer inspection of the dataset, we observed that node 20 (RN) was present in the ED for less than 20 seconds, and made no contact with other nodes during that period. We choose this particular value of α because it was the only one which correctly ranked node 20 last with respect to both BC and RC measures. The time of first appearance of node 20 was set to 4322, the last second possible.

Figure 7.7: Distribution of response variables (Shift 1). On the left we show the distributions of the raw response data; on the right we show the Box-Cox-transformed distributions.

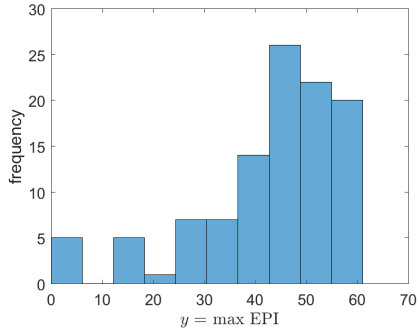
(a) mean EPI



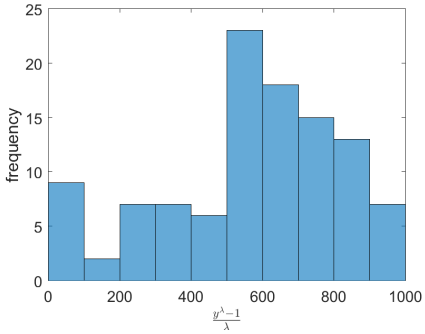
(b) Box-Cox mean EPI ($\lambda = 0.31$)



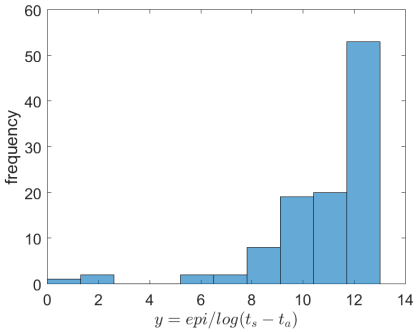
(c) max EPI



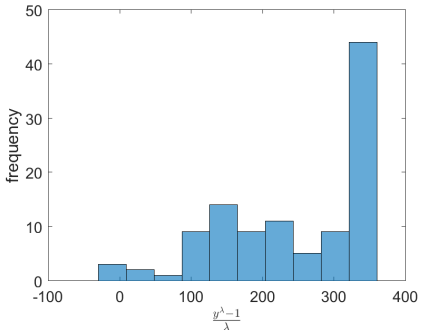
(d) Box-Cox max EPI ($\lambda = 1.81$)



(e) NS-EPI



(f) Box-Cox NS-EPI ($\lambda = 2.68$)



log(BC) score but are associated with a low stochastic mean epidemic size, suggesting that BC measures alone cannot fully explain epidemic outcome.

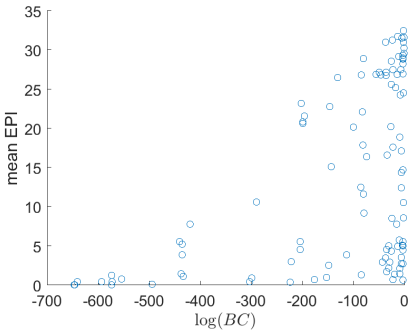
We point out that we also computed BC and RC measures based on this version of dynamic communicability:

$$\hat{Q} = (I + \alpha A^{[1]})(I + \alpha A^{[2]}) \cdots (I + \alpha A^{[M]}).$$

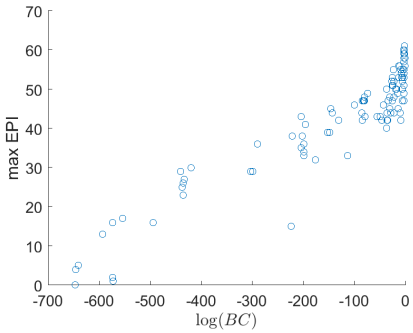
Recall that this imposes the restriction that there is at most one edge per time-step (see Section 3.2.2). This modification only slightly improved the fit of the data (improvements, if any, were in the fourth decimal place), suggesting that in this application, imposing such a restriction makes little difference to BC and RC measures and in particular, has little impact on explaining overall epidemic outcome.

Figure 7.8: Relationship between response and log(BC) (Shift 1)

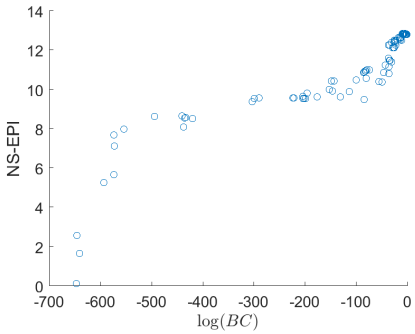
(a) mean EPI ($corr = 0.47$)



(b) max EPI ($corr = 0.93$)



(c) NS-EPI ($corr = 0.92$)



We consider the confounding effects of the duration (D) of activity of the node in the

ED as well as the time (T) at which the node first entered the ED. A variable is said to be a *confounder* if it correlates with both the response and the predictor of interest, thus confounding the true effect of the predictor on the response. Figure 7.9 illustrates that D is indeed a confounder, and should therefore be included in the regression model to account for the effect of D on both log(BC) and epidemic outcome. Figure 7.10 on the other hand, shows that there is a strong multicolinearity effect between T, the time at which a node first enters the ED, and log(BC) (the correlation between them is -0.97). Multicolinearity is known to increase the sensitivity of the parameter estimation, so that coefficient estimates may change erratically in response to small changes in the model or the data. It also causes difficulties in the interpretation of the coefficient estimates: The coefficient associated with log(BC) is usually interpreted as the change in response due to a change in log(BC), assuming that all other variables in the model are kept constant. However, since log(BC) and T are so strongly correlated, it is not possible to keep T constant while increasing log(BC). It is therefore unclear if the change in response is due solely to the increase in log(BC) or due to an accompanying change in T. For these reasons, the variable T will not be included in the regression model.

Table 7.4: Correlation between response and predictors (averaged over 7 shifts)

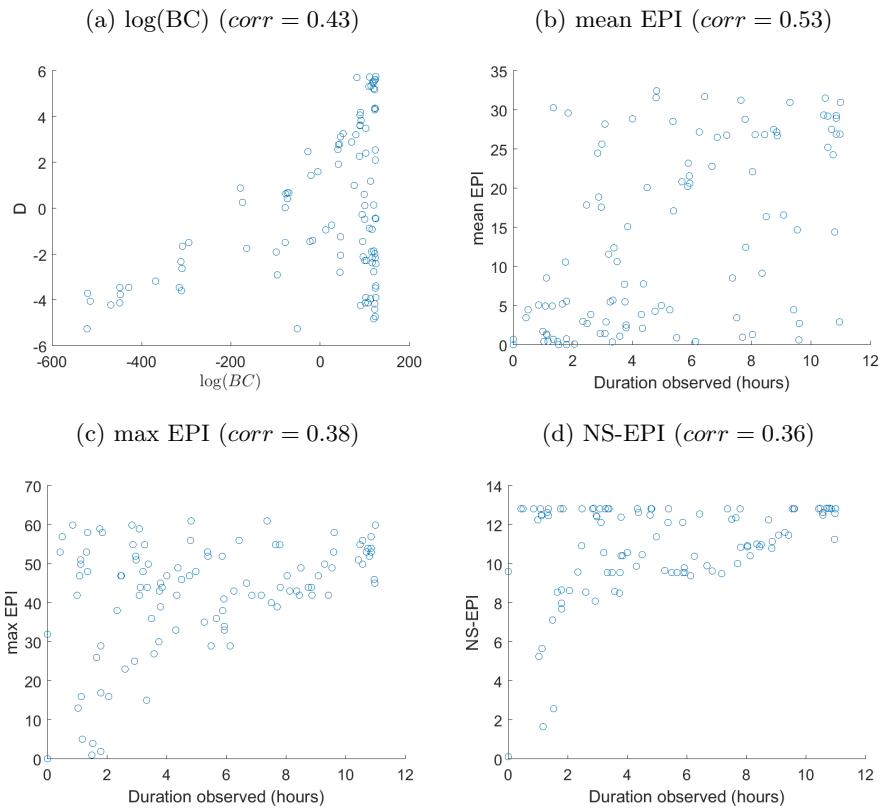
	T	D	log(BC)
mean EPI	-0.43	0.59	0.48
max EPI	-0.85	0.52	0.88
NS-EPI	-0.91	0.47	0.90

Table 7.5: Correlation between log(BC), T and D

	T	D
log(BC)	-0.98	0.54

In Tables 7.4 and 7.5 we report the correlations among the variables, averaged over 7

Figure 7.9: Confounding effect of D (Shift 1)



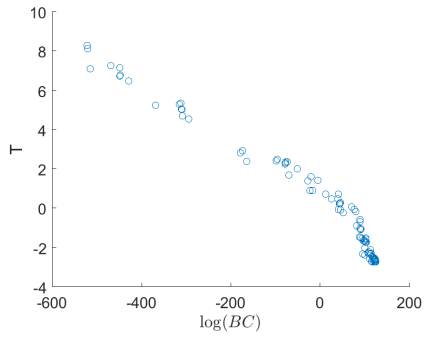
shifts. The behaviors are similar to the discussion based on Shift 1 data. We therefore regress log(BC) on the three different measures of epidemic outcome, including D as a confounder in the model, but not T because of a strong linear relationship with log(BC). Staff/patient category is also included as a binary indicator (see Figure 7.11 for boxplots of log(BC), mean EPI, max EPI and NS-EPI within staff/patient category). For comparison, we consider BD and AD as predictors in lieu of log(BC). Explicitly, we consider models of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 S + \epsilon$$

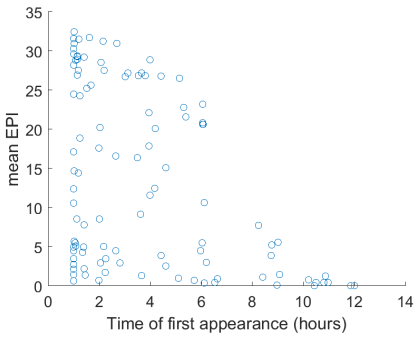
where Y is one of mean EPI, max EPI or NS-EPI, and X is one of log(BC), AD or BD. To have the covariates on a similar scale, we standardize the AD and BD values. All other

Figure 7.10: Confounding effect of T (Shift 1)

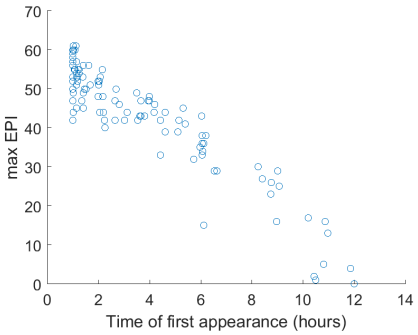
(a) log(BC) ($corr = -0.97$)



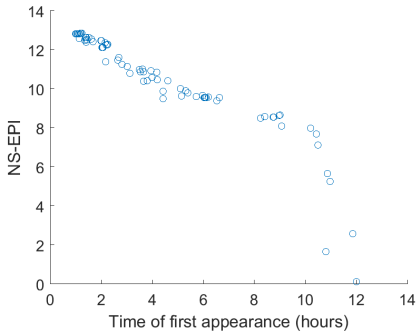
(b) mean EPI ($corr = -0.43$)



(c) max EPI ($corr = -0.93$)



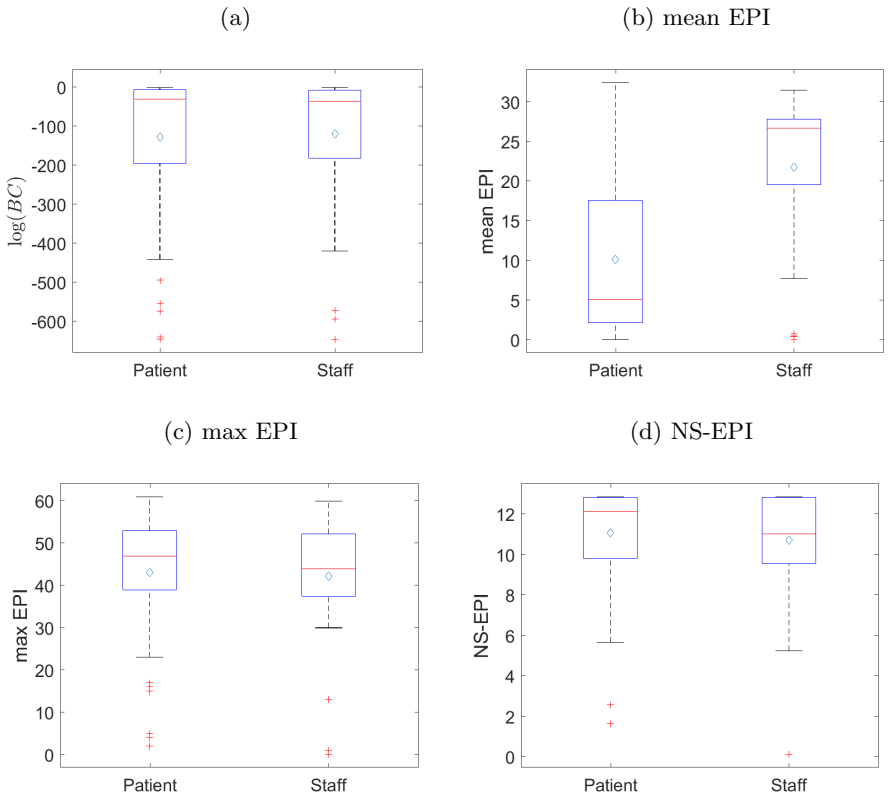
(d) NS-EPI ($corr = -0.94$)



quantitative covariates are mean-centered to assist in the interpretation of the coefficients. We perform standard linear regression (using Matlab's `fitlm`). Model coefficients and associated 95% confidence intervals are reported in Table 7.6.

We report R^2 as a measure of goodness-of-fit. $0 \leq R^2 \leq 1$ can be thought of as a generalized form of squared Pearson correlation when there is more than one predictor. A high value of R^2 is indicative of a good fit to the data, bearing in mind that an increase in the number of predictors *always* increases the R^2 value. For comparison, we report the R^2 values associated single-predictor models of the form $Y = \beta_0 + \beta_1 X$ (in parentheses in Table 7.6). We see that the predictor log(BC) alone is able to explain the variation in the worst-case epidemics (max EPI and NS-EPI), but the low R^2 values associated with the other

Figure 7.11: Boxplots of regression variables within staff/patient category. The distribution of mean EPI is markedly different between the two groups; for the other variables, the distributions are much more similar.



single-predictor models suggest that centrality measures alone cannot always fully account for the variation in epidemic outcome, in which case, the inclusion of D can improve the fit to the data.

Coefficients of the centrality measures are > 0, suggesting that an increase in centrality score is associated with higher epidemic outcome. As expected, there is typically a positive estimated effect of D on epidemic outcome: the longer the seed node is active in the ED,

Table 7.6: Estimated regression coefficients and associated 95% confidence intervals for the eight models under consideration. The response/dependent variable is one of mean EPI, max EPI or NS-EPI associated with the seed node; the predictor of interest is the seed node's centrality measure. Measures studied are log(BC), AD and BD. AD and BD values are standardized; all other quantitative covariates are mean-centered.

	mean EPI			max EPI			NS-EPI		
intercept	10.87*	12.96*	12.55*	43.37*	45.68*	46.11*	11.10*	11.46*	11.51*
	(8.80, 12.93)	(10.66, 15.26)	(10.68, 14.42)	(42.09, 44.66)	(42.39, 48.96)	(43.18, 49.04)	(10.87, 11.32)	(10.90, 12.02)	(11.00, 12.03)
log(BC)	0.02*			0.07*			0.01*		
	(0.01, 0.03)			(0.07, 0.08)			(0.01, 0.01)		
AD (std)		3.91*			0.98			0.09	
		(1.60, 6.22)			(-2.33, 4.28)			(-0.47, 0.65)	
BD (std)			7.12*			5.50*			0.60*
			(5.13, 9.12)			(2.37, 8.62)			(0.05, 1.15)
D (in hours)	0.68*	1.15*	0.22	0.06	2.11*	1.26*	0.00*	0.35*	0.26*
	(0.04, 1.32)	(0.54, 1.75)	(-0.38, 0.82)	(-0.33, 0.46)	(1.24, 2.97)	(0.32, 2.21)	(-0.07, 0.07)	(0.21, 0.50)	(0.09, 0.43)
S	9.21*	2.43	3.75*	-1.7	−9.17*	−10.58*	−0.47*	−1.65*	−1.82*
	(5.12, 13.30)	(-2.56, 7.42)	(0.03, 7.47)	(-4.24, 0.85)	(-16.31, -2.03)	(-16.41, -4.74)	(-0.92, -0.03)	(-2.86, -0.43)	(-2.85, -0.80)
R^2	0.46 (0.23)	0.41 (0.31)	0.56 (0.54)	0.86 (0.86)	0.21 (0.01)	0.29 (0.17)	0.86 (0.85)	0.20 (0.003)	0.23 (0.10)

*$p < 0.05$

R^2 values in parantheses indicate the values obtained in single-predictor models

the larger the epidemic outcome. Observe that since S is a binary indicator,

$$\mathbb{E}(Y|S=1) = (\beta_0 + \beta_3) + \beta_1 X + \beta_2 D$$

$$\mathbb{E}(Y|S=0) = \beta_0 + \beta_1 X + \beta_2 D,$$

since the covariates are mean-centered, the intercept term β_0 is interpreted as the expected epidemic outcome among patients ($S = 0$) with average centrality score ($X = 0$) and average length of duration active in the ED ($D = 0$). On the other hand, β_3 is interpreted as the *difference* in epidemic outcome between the two groups: among staff members with average centrality and D, the expected mean epidemic size is significantly larger (by 9.21 and 3.75) than that associated with patients. (When AD is in the model, the expected difference in mean EPI between the two groups is 2.43, but this result is not significant.) With respect to non-stochastic epidemic measure (NS-EPI), the reverse is true: staff members are associated with slightly smaller NS-EPI measure compared to patients (by 0.47, 1.65 and 1.82), and while these differences are small, they are statistically significant. Similarly, the maximum

epidemic size associated with staff members is on average smaller (by 1.7, 9.17 and 10.58) than max EPI associated with patients.

7.3 Interaction effects

For each of the models in Table 7.6, we report the interaction effects of D and S separately. Explicitly, to examine the interaction effect of D (say), the data is divided into groups based on the quartiles of D, and within each group, the regression of epidemic outcome on centrality is performed, while adjusting for S. This allows us to see how the effect of centrality (β_1) depends on D, thus providing a quantitative grasp on how centrality and D 'interact'. (Analogously for the interaction effect of S on centrality.) In Table 7.7 we see that the coefficients are typically positive, suggesting that regardless of D, an increase in centrality is associated with an increase in epidemic outcome. Strong monotonically decreasing interaction effects are observed for AD: the longer nodes are active in the ED (larger D), the smaller the effect of AD on epidemic outcome. (The same phenomenon is observed for log(BC), albeit on a smaller scale.) This suggests that minimizing AD of nodes who are in the ED for a short amount of time may reduce epidemic outcome more significantly, than if AD is targeted for nodes who have been in the ED for a longer time.

In Table 7.8 we see that the effect of log(BC) is similar among both staff and patients. Observe that AD of patients has a much stronger effect (than AD of staff) on both mean and max EPI. For example, a unit increase of AD among patients increases mean EPI (on average) by 36.4, whereas a unit increase of AD among staff increases mean EPI (on average) by only 2.27. The same is observed for BD: a unit increase in BD (or equivalently, an additional distinct contact) among patients doubles the stochastic epidemic outcomes. However, with respect to the non-stochastic epidemic measure, the effect of centrality is marginal regardless of staff/patient category.

Table 7.7: Interaction by D (in hours) is examined by stratifying the data according to quartiles of D. Within each group, epidemic outcome is regressed on centrality while adjusting for S. The coefficients associated with the centrality measure for each group are reported below.

		$0 < D \leq 2.51$	$2.51 < D \leq 4.78$	$4.78 < D \leq 8.05$	$D > 8.05$
mean EPI	$\log(BC)$	0.01	0.02	0.06*	0.06
		(0.00, 0.03)	(0.00, 0.05)	(0.02, 0.10)	(-0.06, 0.18)
	AD	132.11*	29.56*	4.39	2.06
		(98.85, 165.38)	(7.00, 52.11)	(-5.45, 14.23)	(-0.30, 4.41)
	BD	7.94*	7.05*	11.67*	3.37*
		(2.93, 12.95)	(1.00, 13.10)	(7.34, 16.00)	(0.61, 6.14)
max EPI	$\log(BC)$	0.07*	0.06*	0.09*	0.14*
		(0.06, 0.08)	(0.04, 0.08)	(0.07, 0.10)	(0.08, 0.21)
	AD	96.34	1.46	-0.49	-0.98
		(-52.37, 245.05)	(-30.87, 33.80)	(-9.11, 8.13)	(-2.80, 0.84)
	BD	13.11	3.06	6.50*	1.36
		(-0.10, 26.33)	(-5.30, 11.42)	(1.53, 11.48)	(-0.88, 3.60)
NS-EPI	$\log(BC)$	0.01*	0.01*	0.01*	0.03*
		(0.01, 0.01)	(0.01, 0.01)	(0.01, 0.01)	(0.02, 0.04)
	AD	11.58	-3.24	-0.42	-0.12
		(-16.09, 39.26)	(-7.88, 1.41)	(-1.61, 0.77)	(-0.41, 0.18)
	BD	1.79	-0.67	0.54	0.24
		(-0.71, 4.29)	(-1.91, 0.56)	(-0.22, 1.30)	(-0.12, 0.60)

*$p < 0.05$

7.4 Prediction

We train the regression models in Section 7.2 on Shift 1 data (training set) and use the estimated regression coefficients to predict epidemic outcomes for Shifts 2 to 7 (testing data)[2]. The epidemic outcomes (mean/max EPI, NS-EPI) are considered to be the ground-truth. Predictions are performed separately for each shift. Predicted values ($\hat{Y}$) are compared to the values obtained by simulation (Y). Prediction errors are computed in two ways (where

[2]To reduce variability, one can also cross-validate with each shift separately as the training data set, and average the prediction results over the rounds. This will be left for future work.

Table 7.8: Data is stratified according to staff/patient category. Within each group, we regress epidemic outcome with respect to the centrality measure, while adjusting for D. The coefficients associated with the centrality measure for each group are reported below.

		staff	patient
mean EPI	$\log(BC)$	0.03^*	0.02^*
		(0.01, 0.05)	(0.01, 0.03)
	AD	2.27^*	36.4^*
		(0.72, 3.81)	(25.40, 47.40)
	BD	1.57	10.05^*
		(-1.38, 4.52)	(7.42, 12.67)
max EPI	$\log(BC)$	0.06^*	0.07^*
		(0.05, 0.08)	(0.06, 0.08)
	AD	-0.26	12.76
		(-2.51, 1.99)	(-5.94, 31.46)
	BD	2.15	5.76^*
		(-1.61, 5.91)	(1.16, 10.37)
NS-EPI	$\log(BC)$	0.01^*	0.01^*
		(0.01, 0.02)	(0.01, 0.01)
	AD	-0.08	0.27
		(-0.52, 0.37)	(-2.81, 3.36)
	BD	0.30	0.35
		(-0.45, 1.06)	(-0.43, 1.12)

$^*p < 0.05$

n is the total number of nodes):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2}$$

$$\text{BIAS} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right).$$

Since the range of the epidemic outcomes differ by shift, we standardize RMSE and BIAS values by multiplication by $100/n$. The average of the standardized RMSE and BIAS values over 6 shifts are reported in Table 7.9. A standardized RMSE value of 10.39 means that,

on average, the predicted epidemic outcome (mean/max EPI) fails to correctly capture the observed outcome by a factor of 10.39% of the total population under study. The sign of the bias provides an indication of over/under-estimation: a positive bias of 1.45 means that on average, the predicted outcome overestimates the observed outcome by 1.45% of the population. From the perspective of predicting epidemic outcome, a positive bias is preferable to a negative bias, since one would prefer to err on the side of caution. From Table 7.9, we see that for all the models under consideration, predicted values overestimate the observed outcome. In addition, predictions of the mean epidemic size are consistently more accurate than predictions of the maximum epidemic size, while predictions of the non-stochastic epidemic measure are the most accurate of all.

To assess the impact of including network centralities in developing predictive models, we consider models without centrality: that is, we regress the response against D and S only. For comparison, we also consider models with T in lieu of centrality. The difference in prediction error relative to the null model quantifies the change in predictive power: a *decrease* in RMSE/BIAS values indicates a stronger predictive model compared to the null model. The results in Table 7.9 show that inclusion of centrality almost always reduces RMSE, while BIAS sometimes increases, albeit only slightly. This means that the inclusion of centrality results in predictions that tend to slightly increase overestimation of the true values, but the predictions are overall more accurate.

Figure 7.12 displays the predictions '+' relative to the observed epidemic outcomes 'o' for Shift 5. (For other shifts see Appendix H.) For comparison, we plot the predictions based on the null model where T is used in place of centrality. The full model generates predictions that capture the overall trend of the data very well, and there is significant improvement compared to the predicted trends based on single-predictor models (not shown). Observe that the null model is also effective in capturing the data trend, suggesting that temporal markers such as D (duration observed) and T (time of first appearance) play an important

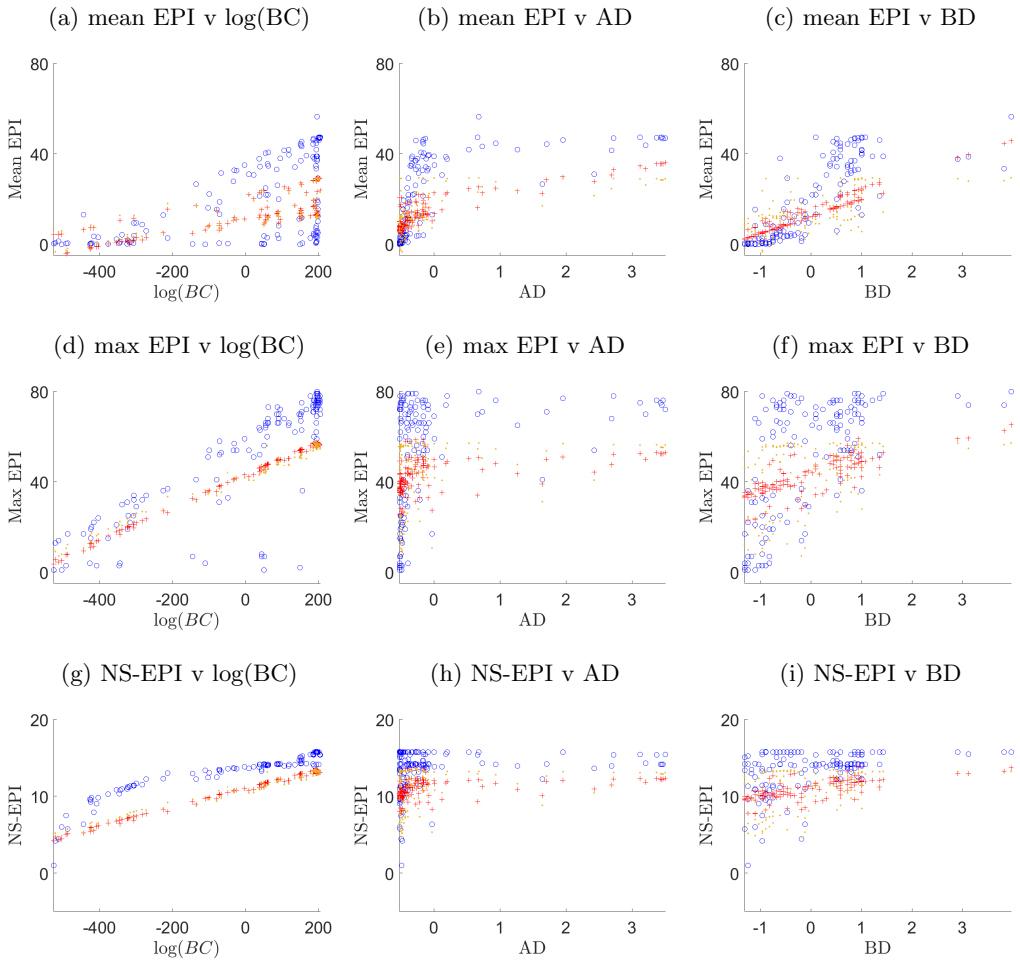
Table 7.9: Prediction errors for full models. Reported values are standardized and averaged over 6 shifts. For comparison, prediction errors based on the null models (the first, where T is used in lieu of centrality, and the second, where the response is regressed against D and S only) are also reported.

Response	Predictors			RMSE	BIAS
mean EPI	log(BC)	D	S	10.39	1.45
	AD	D	S	10.07	1.56
	BD	D	S	9.56	1.54
	T	D	S	10.53	1.44
		D	S	10.64	1.48
max EPI	log(BC)	D	S	15.40	4.28
	AD	D	S	19.68	4.40
	BD	D	S	19.11	4.43
	T	D	S	15.58	4.26
		D	S	19.72	4.38
NS-EPI	log(BC)	D	S	2.01	0.02
	AD	D	S	2.75	0.04
	BD	D	S	2.72	0.04
	T	D	S	1.94	0.01
		D	S	2.75	0.03

role in reproducing realistic prediction patterns.

A key observation is that the inclusion of the temporal measure log(BC) improves predictive power of worst-case epidemics (max EPI and NS-EPI) more than the aggregated/non-temporal measures AD and BD, but with respect to mean epidemic size, inclusion of BD reduces RMSE by the largest margin. This may be due to the fact that BD has the strongest linear relationship with mean epidemic size (see R^2 values in parentheses in Table 7.6) compared to the other centrality measures. We also observe that the inclusion of T in lieu of log(BC) has the same effect on predictive power, which is not surprising because of the high correlation between the two variables. Overall reduction in RMSE suggests that inclusion of network-based centrality does improve predictive power, but the improvement is marginal, and whether or not this level of improvement is worth the computational effort remains

Figure 7.12: Predictions based on full models (Shift 5). Observed values are denoted 'o' (blue) and predicted values based on full models are denoted '+' (red). Predictions based on the null model: $Y \sim T + D + S$ are denoted '·' (yellow). For remaining shifts see Appendix H.



subjective, especially since knowledge of T (the earliest time at which nodes entered the ED) is sufficient to replicate the results based on log(BC).

Chapter 8

Targeted edge manipulation

In this chapter, we consider an alternative approach to quantify the relationship between contagion and centrality. The goal is to use network centrality to strategically modify the edge-structure of the temporal network. We seek to study the corresponding changes in both network structure, as well as epidemic outcome. We are motivated by previous work in [1, 4, 5], which has shown that using centrality-based strategies to target nodes and/or connections can have a strong effect on the network. In [4, 5], analysis was performed on static networks. We modify and adapt their methods to answer this question in the temporal context: if connections between highly central nodes are removed, what effect does this have on both the network, and the associated epidemic outcome?

We consider the epidemic effect based on the deterministic approach described in Chapter 6. Recall that the deterministic infection process provides a way to tune infection rates based solely on the observed, contiguous contact time between nodes. We consider contagion processes where infection occurs after 10, 15, 20, 25 minutes of contiguous contact time, and will refer to these as different 'infection strategies'. In addition, we would like to examine if, and how, the relationship between temporal centrality and epidemic effect depends on the infection strategy employed. Such analysis will provide a more concrete handle on

the relationship between temporal centrality and epidemic outcome, and may provide some insight into the types of situations in which temporal centrality is more applicable or effective as an explanatory or predictive factor of epidemic effect.

By strategically eliminating connections between highly central nodes, the changes in observed epidemic outcome will provide some insight into how network structure influences epidemic spread. Our analysis is unique in the way that it incorporates temporal information based on observed data, and we hope that our work will motivate further developments of such methods.

8.1 Epidemic measures

As in Chapter 6, we want to associate an epidemic measure associated with each node in the network. Consequently, the infection process is seeded by only one initial source, cycling over all nodes in the network. Per infection strategy, we compute and analyze a number of epidemic measures associated with each initial source.

A natural epidemic outcome associated with each initial source i is the final epidemic size (EPI_i), counted at the end of the shift. In order to incorporate some temporal information regarding *when* the infections took place, we also consider a measure of the form

$$W_i = \frac{\text{EPI}_i}{\sum_j T_{ij}}$$

where T_{ij} is the time at which node j became infected, when node i is the initial source of infection. (Note that node i, while being the initial source of infection, does not necessarily infect node j directly.) If, by the end of the shift, node j was not infected, we set $T_{ij} = 0$. The rationale for the measure W_i is this: an initial source that is able to infect a large number of nodes in a short amount of time should be associated with a larger epidemic measure, than an initial source that either requires a longer time to infect the same number

of nodes, or an initial source that is able to infect fewer nodes overall. We also rank nodes according to the *difference* in W_i:

$$W_{i,\text{diff}} = W_{i,\text{orig}} - W_{i,\text{new}} = \frac{\text{EPI}_{i,\text{orig}}}{\sum_j T_{ij,\text{orig}}} - \frac{\text{EPI}_{i,\text{new}}}{\sum_j T_{ij,\text{new}}}$$

where $W_{i,\text{orig}}$ is the value of W_i associated with the original network, and $W_{i,\text{new}}$ is the value of W_i associated with the edge-manipulated network. The rankings based on $W_{i,\text{diff}}$ are compared with the rankings based on centralities computed on the original network. Only nodes that have non-zero epidemic outcome are considered for ranking purposes. Specifically, nodes i for which $\text{EPI}_{i,\text{orig}} = 0$, and nodes for which $W_{i,\text{orig}} = W_{i,\text{new}}$ are discounted in the ranking process. Furthermore, if $\text{EPI}_{i,\text{new}} = 0$, we set $W_{i,\text{new}} = 0$ to ensure that $W_{i,\text{diff}}$ is largest possible.

Finally, we look at the number of initial sources able to infect $5, 10, 20, 30, 40\%$ of the network, where the ability to infect the respective percentage of the network is referred to as the 'potency' of the node. We also compute the average time taken to infect 5, 10, 20, 30, 40% of the network, where the average is taken over nodes which achieved the respective level of potency.

8.2 Edge manipulation

In this section we discuss different edge manipulation strategies employed based on dynamic communicability. Motivated by the work in [1, 4, 5], we manipulate the network by targeting highly central nodes. Our initial strategy is to target interactions between high broadcasters and high receivers, as this has been shown in [4] to be a highly effective way to influence the overall connectivity of the network. In [4], the authors first assign edge centrality measures in the following way: for every directed edge $a \to b$, multiply the broadcast score of a with the receive score of b. Directed edges are then eliminated based on their assigned scores.

In the temporal setting, there is no immediate analogous method to assign centrality scores to edges since the nodes' broadcast and receive scores depend very much on the histories of dynamic walks, and themselves change over time. The temporal dimension also requires the added consideration of *when* edge manipulation should take place. As a first approach, we apply the conditions based on temporal centrality determined at the final time step.

We present the results based on Shift 7 data. Recall from Chapter 5 the following notation: Let top10BC denote the set of top 10 nodes ranked according to broadcast centrality BC, and analogously for top10RC. Consider

$$X = \text{top10BC} \setminus \text{top10RC}$$

$$Y = \text{top10RC} \setminus \text{top10BC}$$

$$Z = \text{top10BC} \cap \text{top10RC}.$$

Based on Shift 7 data, we have $X = \{3, 12, 23, 28, 33, 92\}$, $Y = \{11, 17, 24, 30, 32, 35\}$ and $Z = \{4, 5, 7, 8\}$ (not in order of centrality). There are 16 distinct nodes in $X \cup Y \cup Z$, out of $n = 126$ nodes in total. Note that there is only one patient present among these top-ranking nodes: node 92. It is also worth pointing out that nodes in Z (high broadcasters and high receivers) are staff members not including RN's (for example, adminstrative staff), while all nodes in Y apart from node 11 are RN's.

The goal of this work is to see if edge-deletion strategies based on temporal network centrality have a significant impact on epidemic outcome. Previous work [58, 82] has shown that targeting inter-community links is more effective in containing epidemic spread than targeting intra-community links. Intuitively, nodes within tightly-knit communities are connected by many different paths/walks; consequently, removing only a subset of connections within such a group can at best slow down the epidemic spread, but will not have a strong impact on the overall epidemic effect. The notion of a community in the temporal setting

is difficult to define. However, as discussed in Chapter 5, it can be argued that the nodes in $Z = \text{top10BC} \cap \text{top10RC} = \{4, 5, 7, 8\}$ form a 'temporal community' since Z nodes interact a lot with each other across both space and time. Recall Figure 5.1a, which shows that the interactions between the nodes in Z (aggregated over time) are of a similar magnitude, while Figure 5.1b shows that all four Z nodes are simultaneously in the same location for much of the shift. Consequently, we conclude that over time, the nodes in Z interact as a tightly-knit group, with no particular node dominating the interactions.

It is therefore reasonable to consider Z as a 'community' of nodes in this network. Our first edge-deletion strategy (denoted 'deleteXY') leaves Z nodes alone, while targeting only interactions between nodes in X and nodes in Y. Note that interactions among nodes in X remain unchanged and similarly for nodes in Y – only interactions of the form (a, b) where $a \in X$ and $b \in Y$ are deleted. (Based on how the edge information is stored, we also eliminate edges of the form (a, b) where $a \in Y$ and $b \in X$.) The next strategy, denoted 'deleteXYZ', removes XZ and YZ interactions in addition to XY interactions. Again, node interactions within each group remained unchanged. For comparison, we also looked at what happens when *all* interactions associated with each group are deleted (denoted 'deleteXall', 'deleteYall' and 'deleteZall'), as well as what happens when all interactions associated with each *pair* of groups are deleted (denoted 'deleteXYall', 'deleteXZall' and 'deleteYZall'). Finally, we consider deleting all X, Y and Z interactions ('deleteXYZall'). These latter strategies are the same as effectively removing the respective groups from the network. We emphasize that this is merely an experiment to help us better understand the role of network-based centrality in the context of epidemic spread – we are not advocating that such drastic measures (akin to banning/removing people from the ED) should be applied in practice. The percentage of edges removed based on each of these strategies are shown in Table 8.1. Edge deletions were first performed on the raw 1-second edges, before aggregating into 10-second time-frames as usual. There was little difference in the

percentages (of edges deleted) based on 1-second edges compared to 10-second edges: the differences occurred in the third or fourth significant figure. Consequently, in Table 8.1, we do not distinguish between them, referring only to 'temporal edges'. For reference, the total number of raw 1-second edges, 10-second aggregated edges, and binarized edges[1] of the original network are shown in Table 8.2.

Table 8.1: Edge-deletion strategies (Shift 7). Edge deletions were performed on raw 1-second edges before aggregating into 10-second time-frames. There was little difference between the percentages of edges deleted based on 1-second edges compared to 10-second edges. For brevity, such edges are referred to as temporal edges. Binarized edges are those based on the binarized, aggregated network[2].

		% deleted edges	
Strategy	Description	temporal	binarized
deleteXY*	delete XY edges	7	2
deleteXYZ*	delete XY, XZ and YZ edges	30	5
deleteXall**	delete all X interactions	22	12
deleteYall**	delete all Y interactions	44	14
deleteZall**	delete all Z interactions	36	13
deleteXYall**	delete all X and Y interactions	58	24
deleteXZall**	delete all X and Z interactions	53	23
deleteYZall**	delete all Y and Z interactions	61	25
deleteXYZall**	delete all X, Y and Z interactions	71	33

*interactions among nodes within each group remain

**effectively removes group(s) from the network

From Table 8.2 we see that the network is fairly sparse: the binarized network has density $1290/\binom{126}{2} = 16\%$. The results in Table 8.1 are also interesting in their own right: more than

[1]Recall that binarized edges are edges based on the binarized, aggregated matrix. The presence of an edge (i, j) in the binarized matrix means that nodes i and j shared a location at some point over the entire shift. All temporal information is lost. The degree based on the binarized matrix essentially counts the number of distinct contacts made over the entire shift.

[2]See Footnote 1.

Table 8.2: Number of original edges (Shift 7)

1-sec	10-sec	binarized
2,703,591	277,586	1,290

70% of temporal edges are attributed to only 16 nodes (in $X \cup Y \cup Z$) identified by dynamic communicability as high broadcasters and high receivers. Among these edges, slightly less than half (30%/70%) consist of interactions within these 16 nodes. Furthermore, nodes that are both high broadcasters and high receivers (Z) are responsible for the majority of the interaction between the groups X and Y: out of 30% of inter-group interactions, a large proportion (24%/30%) occur with Z. We also observe that nodes in X (high broadcasters but not high receivers) account for the least number of temporal edges (22%) compared to the other groups, reinforcing the idea that broadcasting ability depends more on the time-sensitive quality of links rather than sheer quantity.

8.3 Effect of edge manipulation on dynamic total communicability (DTC)

In a similar vein to the methods in [4, 5], we propose summing up the BC/RC scores over all nodes as a global measure associated with the network, which indicates how well communication takes place via time-respecting walks. We call this measure *dynamic total communicability*, abbreviated by DTC. Explicitly, we define DTC as the sum of all entries of Q, where Q is the dynamic communicability matrix described in Section 3.2.

$$DTC = \mathbb{1}^T \cdot Q \cdot \mathbb{1} = \mathbb{1}^T \cdot Q^T \cdot \mathbb{1}.$$

Note that DTC can be viewed as either the sum of BC values over all nodes ($\mathbb{1}^T \cdot Q \cdot \mathbb{1}$), or the sum of RC values over all nodes ($\mathbb{1}^T \cdot Q^T \cdot \mathbb{1}$). DTC can be considered a global,

structure-based measure associated with the temporal network: A larger value of DTC (on a network with the same number of nodes) would suggest that there are a greater number of 'short' walks connecting nodes in the network, and therefore that the network has the potential, at a structural level, to be more effective at overall communication.

In the temporal setting, additional care must be taken in computing DTC. Recall that in the typical computation of Q, normalization is performed at each time step to ensure that matrix overflow does not occur. Normalization does not affect the rankings of the nodes obtained, since all nodes are subject to the same normalizing factor at each time-step. (Indeed, the discussion in Section 4.3 provides strong evidence that the node rankings are correct.) However, since the normalization factors are network-dependent, comparisons of Q across different networks poses some challenges. Entries of Q also depend on the numeric choice of parameter α, therefore, in order to make meaningful comparisons of how DTC changes after edge manipulation, we must **a)** use the same value of α on both networks, and **b)** use the same normalizing factor at each time-step if possible. Let $\rho_0^t = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A^{[t]}\}$ be the spectral radius of the adjacency matrix of the original network at time-step t, and let ρ_1^t be the corresponding quantity for the manipulated network. It is a well-known fact in Perron-Frobenius theory that the spectral radius of a non-negative matrix A is a monotonic function of the entries in A. Therefore, since any edge manipulation can only reduce the number of entries of the adjacency matrix at any time step, $\max_t \rho_1^t \leq \max_t \rho_0^t$, and consequently,

$$\frac{1}{\max_t \rho_0^t} \leq \frac{1}{\max_t \rho_1^t}.$$

The left-hand side of the above equation is the upper bound for α associated with the manipulated network, while the right-hand side is the upper bound for α associated with the original network. To deal with **a)**, we use the smaller of the two quantities: for both

the original and manipulated networks, choose[3]

$$\alpha = 0.25 * \frac{1}{\max_t \rho_0^t}.$$

There are a number of ways to deal with **b)**. At each time step, we can compute $(I - \alpha A^{[t]})^{-1} \cdot \mathbb{1}$ for both the original and manipulated networks simultaneously, choosing either the larger or smaller of the two norms as the common normalizing factor. For some edge manipulation strategies, this resulted in values that differed by more than 100 orders of magnitude, rendering comparisons across such values quite meaningless. Another approach is to compute unnormalized values of $\tilde{Q} \cdot \mathbb{1}$ over sliding windows of length $w << T$, then taking the average over all windows. We consider windows of length $w = 50, 100, 150, \ldots, 500$. Explicitly, for each window length w, we perform the following computation (where T is the total number of time steps):

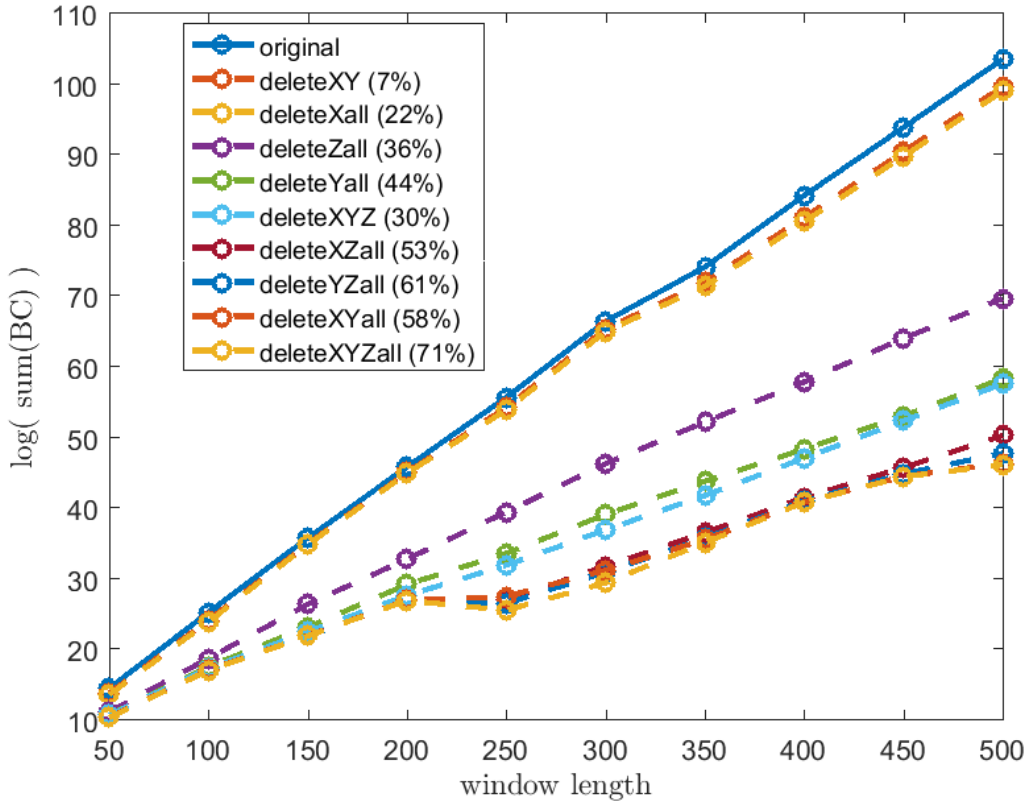
```
avg = 0;
w = window_length;
num_windows = T - w + 1;
for t = 1 : num_windows
        x = DTC computed over time steps t : t + w;
        avg = avg + x/num_windows;
end
```

We divide by `num_windows` at each time step in order to avoid overflow. Also, the final window consists of the last $w - 1$ matrices.

In Figure 8.1, per edge-strategy we display the average value of DTC over sliding windows of varying lengths. We emphasize that our methods here provide some measure of the effect

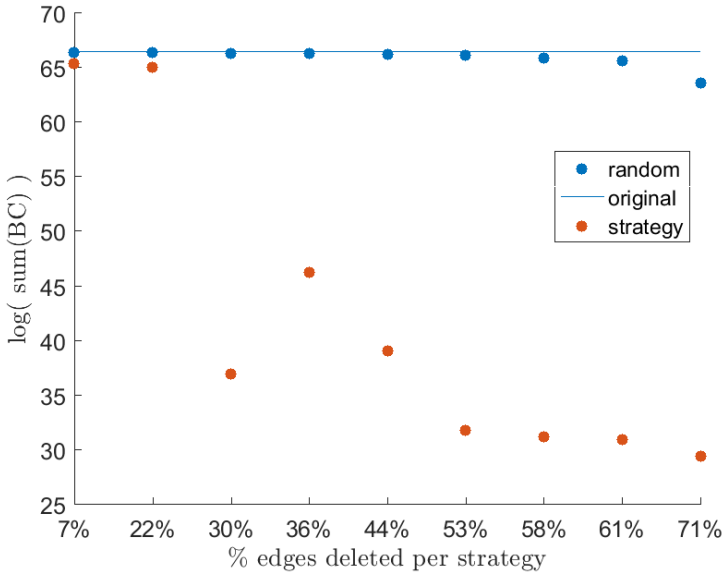
[3]In line with earlier work, we work with $\alpha = 0.25 * \alpha_{\max}$. Recall the findings in Chapter 4, in particular, that node-rankings based on dynamic communicability are relatively robust with respect to the choice of α in the regime $[0.25, 0.85] * \alpha_{\max}$.

Figure 8.1: Effect of edge manipulation on dynamic total communicability (DTC). We show the results computed by averaging the sum of entries of Q over sliding windows of various lengths. For interpretive ease, legend entries are shown in decreasing order of DTC when window length is 500.



of edge manipulation on the structure of the network – the *difference* in values of DTC that arise from edge deletion is of more interest than the numerical values themselves. Indeed, the longer the time window, the better the approximation of the true effect of edge manipulation on DTC. Figure 8.1 shows that as window length w increases, the reduction in DTC resulting from edge deletion grows exponentially, providing strong evidence that edge manipulation strategies based on dynamic communicability can have a significant impact on underlying structural characteristics of the network. It is noteworthy that the strategy deleteXYZ, which removed only interactions *between* the groups X, Y and Z of the 16 key

Figure 8.2: Effect of random deletion on dynamic total communicability (DTC). Window length = 300. Per edge strategy, we compute DTC after randomly deleting a percentage of temporal edges as shown in Table 8.1. We plot the minimum over 5 iterations. For comparison, we plot the value of DTC of the original network, as well as values obtained after strategic edge-deletion.



players identified by dynamic communicability, resulted in a disproportionate reduction in DTC: this strategy deletes only 30% of temporal edges but outperforms other strategies in reducing DTC, such as deleteZall and deleteYall which removed a greater proportion of edges (36% and 44% respectively). In other words, strategically removing connections between the groups X, Y and Z has a stronger effect on DTC than removing each group entirely from the network. This result provides a strong indication that the edges between high broadcasters and high receivers are in some sense special, and have a particularly strong effect on the global, structural measure of connectivity as determined by DTC.

In Figure 8.2 we show the effect of random deletion on DTC, relative to the effect of strategic edge-deletion methods. For every edge-deletion strategy proposed, we remove the same percentage of edges at random. We use window length $w = 300$, and plot the minimum

over 5 iterations. See Table 8.1 for the strategies associated with the percentages listed on the horizontal axis. It is evident that the majority of the proposed edge-deletion strategies significantly outperform random deletions with respect to reducing DTC. Strategy deleteXY (removes 7%) and strategy deleteXall (removes 22%) showed the least improvement relative to randomly deleting the same percentage of edges. We see a dramatic improvement from random deletions when interactions between X, Y and Z are removed: deleteXYZ (30%) again outperforms deleteZall (36%) and deleteYall (44%). It is arguable that, relative to the percentage of edges removed, strategy deleteXYZ, which involves only inter-group interactions, has the strongest effect on DTC. This result, together with the discussion of Table 8.1, provide an interesting insight: interactions involving Z nodes (nodes which rank highly as both broadcasters and receivers) have a disproportionate effect on the network. When temporal dynamics are in play, it is perhaps worthwhile to combine the dual notions of broadcasting and receiving in determining the 'importance' of a node.

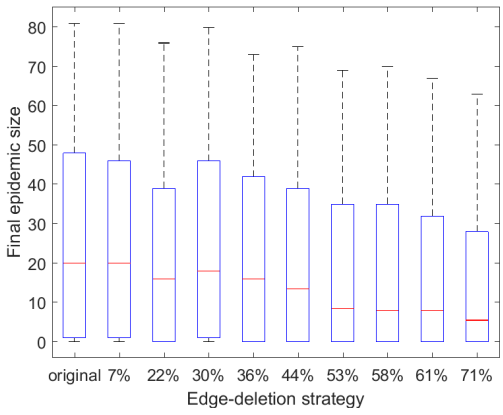
8.4 Effect of edge manipulation on epidemic outcome

We turn our attention to the epidemic effect associated with the centrality-based edge manipulation strategies. In Figure 8.3, we compare the distributions of final epidemic size (over all initial sources) associated with each edge-deletion strategy. Effect on final epidemic size is minimal when infection occurs after 10 minutes of contact time, but increases as longer contact time is required to spread the infection. We see that when infection requires more than 10 minutes of contact time to spread, the edge-strategy deleteXYZ (30%) which targets interactions between top-ranked nodes only, has the same effect on final epidemic size as strategies deleteZall (36%) and deleteYall (44%), suggesting that the targeted removal of connections between highly-ranked nodes can have a relatively larger impact on final epidemic size than removing the same nodes entirely from the network.

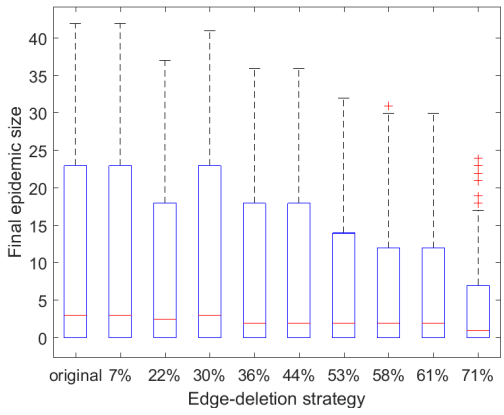
As discussed in Section 8.1, we associate with each node an epidemic measure of the

Figure 8.3: Per infection strategy, compare the distribution of final epidemic size (over all initial sources) associated with each edge-deletion strategy, relative to the original network. Edge-deletion strategies are labeled by the percentage of edges deleted as shown in Table 8.1. On each box, the horizontal line is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

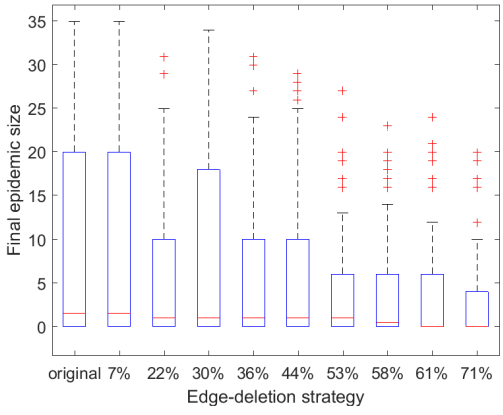
(a) 10 min



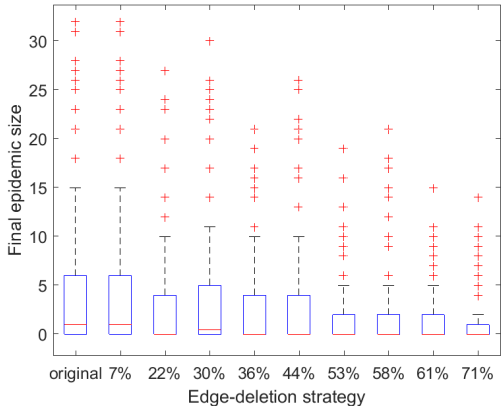
(b) 15 min



(c) 20 min



(d) 25 min



form $W = \text{EPI}/\sum T$. In Figure 8.4 we compare the distributions of W over all initial sources, and over all edge-deletion strategies. The overall change in epidemic effect post edge-deletion remains minimal, and we see the presence of outliers (regardless of infection strategy) which are not eliminated even after the removal of over 70% of the edges. This

is surprising, and suggests that the network is, in some sense, very robust with respect to contagion. It is likely that the presence of many time-respecting walks allows for contagion to spread even if a large percentage of edges are removed. It is interesting to note that the 25th and 75th percentiles of the distributions do not change much (relative to infection on the original network), but the medians are driven down when infection takes at least 20 minutes to spread, and when a large proportion of edges are removed.

In Figure 8.5, we plot the rank correlations between the rankings based on

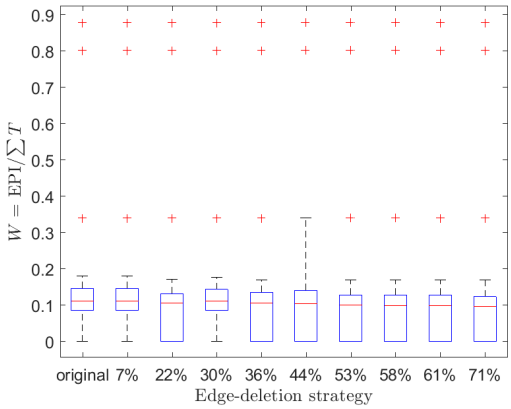
$$W_{i,\text{diff}} = W_{i,\text{orig}} - W_{i,\text{new}} = \frac{\text{EPI}_{i,\text{orig}}}{\sum_j T_{ij,\text{orig}}} - \frac{\text{EPI}_{i,\text{new}}}{\sum_j T_{ij,\text{new}}},$$

and the centrality measures, BC, RC, BD and AD, computed on the original network. For interpretive ease, legend labels are in decreasing order based on BC rankings. We observe that overall, BC rankings correlate the strongest with the epidemic-based ranking, suggesting that nodes with high BC score in the original network tend to be associated with a larger decrease in epidemic outcome (as measured by W_{diff}) when edges are strategically removed from the network. Recall that Y consists of nodes that rank highly as receivers but do not rank highly as broadcasters. With respect to the edge strategy deleteYall (44%), BC rankings based on the original network correlate poorly with epidemic rankings, but interestingly, the converse is true for RC rankings, suggesting that removing Y nodes from the network serves to strengthen the relationship between the receive scores of the remaining nodes and their associated measure of virulence.

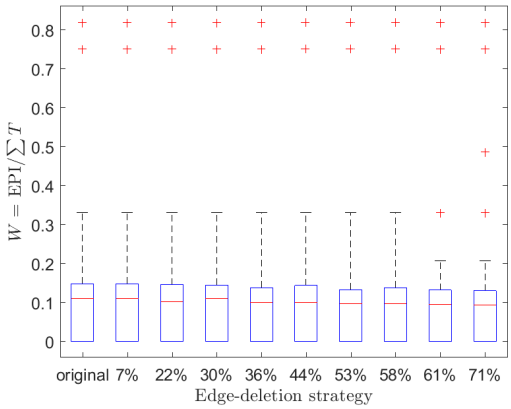
We define 'potent' nodes as nodes which are able to infect a certain percentage of the network. Over different levels of potency (5, 10, 20, 30, 40%), we plot in Figure 8.6 the reduction in the number of potent nodes after edge deletion. For interpretative ease, the legend labels are in decreasing order of the maximum per edge-strategy. Unsurprisingly, the strategy deleteXYZall (71%) resulted in the largest decrease in the number of potent nodes. It is interesting to note that when infection occurs after 20 minutes of contact time,

Figure 8.4: Per infection strategy, compare the distribution of $W = \text{EPI}/\sum T$, which tempers final epidemic size by infection time. Edge-deletion strategies are labeled by the percentage of edges deleted as shown in Table 8.1. On each box, the horizontal line is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

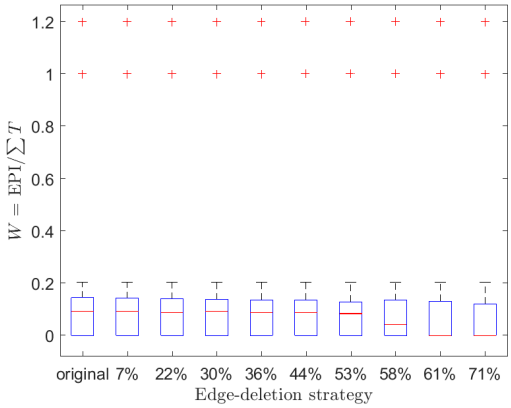
(a) 10 min



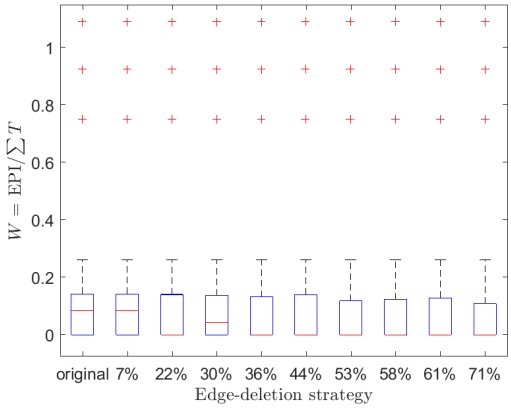
(b) 15 min



(c) 20 min



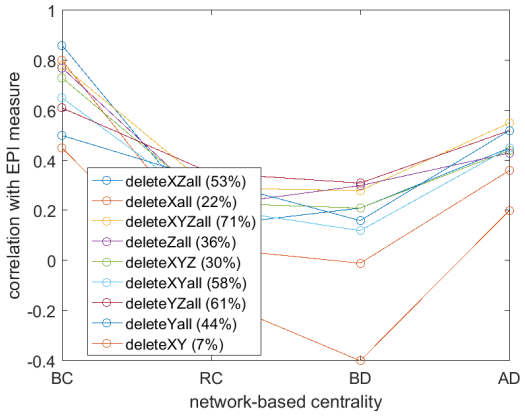
(d) 25 min



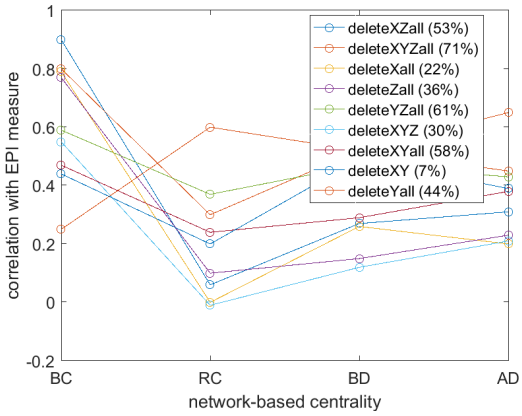
no nodes were able to achieve more than 30% potency. With respect to this measure, the performance of the edge strategies appears to be correlated with the percentage of edges deleted: strategies which delete a larger percentage of edges tend to be associated with a larger reduction in the number of potent nodes. It is interesting to note that the strategy deleteXall (22%) consistently outperforms the edge strategy 'deleteXYZ' (30%) regardless

Figure 8.5: Rank correlations between an epidemic measure W_{diff} and network-based centrality measures. For interpretive ease, legend labels are in decreasing order based on BC rankings.

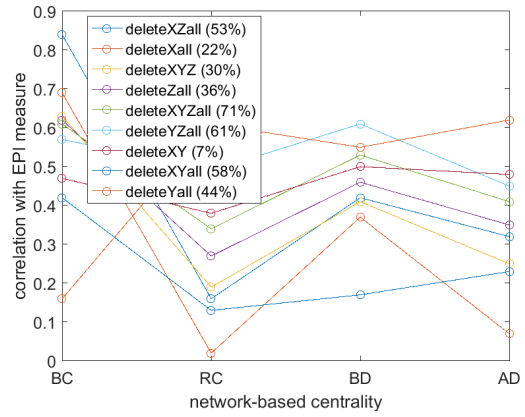
(a) 10 min



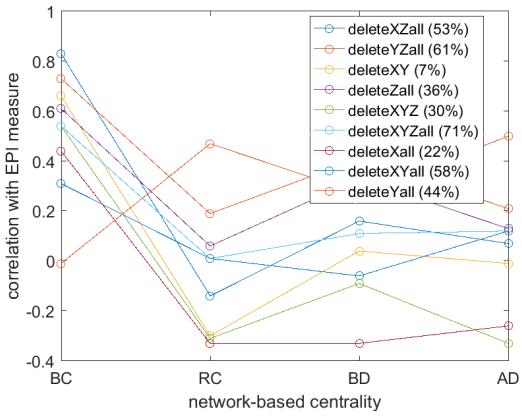
(b) 15 min



(c) 20 min



(d) 25 min

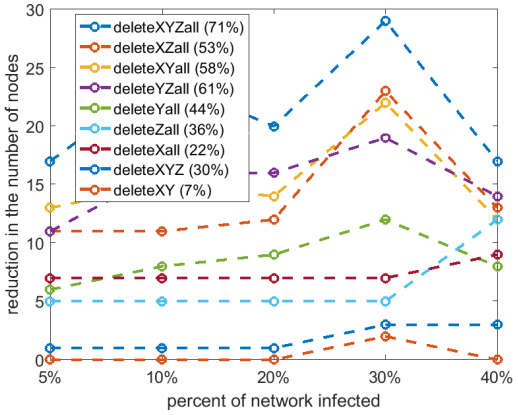


of how long it takes to spread the infection.

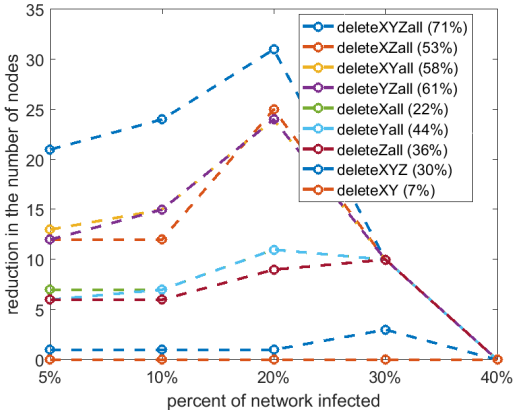
In Figure 8.7 we plot the average time taken to infect a percentage of the network (5, 10, 20, 30, 40%), where the average is computed over the number of nodes which achieved the respective level of potency. We plot the difference in average time taken (in hours), before and after edge deletion. A negative value means that edge-deletion resulted in a shorter time (on average) to reach the level of potency. This tends to happen for higher levels of

Figure 8.6: Potent nodes are nodes which are able to infect a certain percentage of the network. Over different levels of potency (5, 10, 20, 30, 40 %), we plot the reduction in the number of potent nodes after edge deletion. For interpretative ease, legend labels are in decreasing order of the maximum over all potency levels.

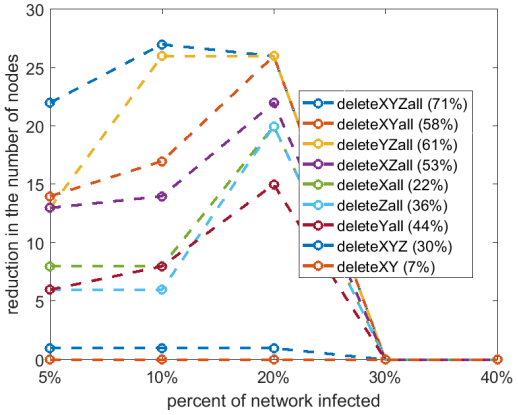
(a) 10 min



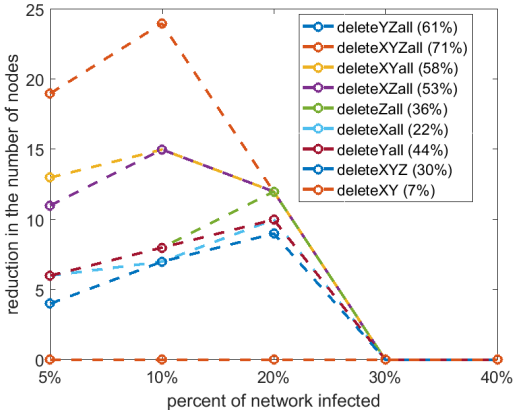
(b) 15 min



(c) 20 min



(d) 25 min

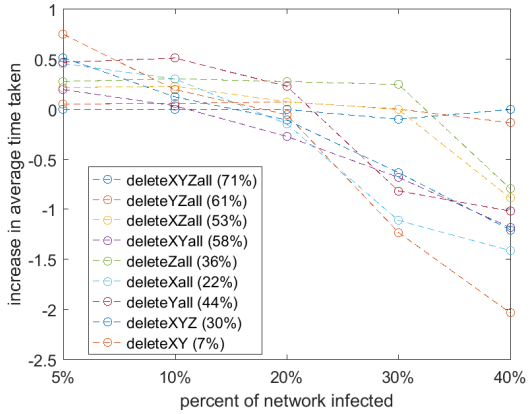


potency, because in the edge-manipulated network, fewer nodes were able to infect such a large percentage of the network, driving down the average value. For interpretive ease, legend labels are in decreasing order based on 5% potency. The increase in average time taken to infect 5% of the network is at most 45-50 minutes. Again, we see that the larger the percentage of edges deleted, the longer it takes to infect 5% of the network, with the

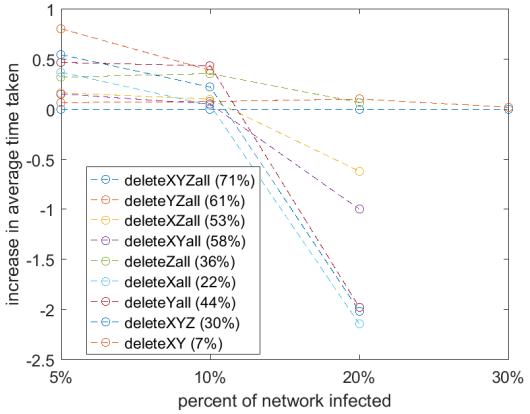
exception of deleteXall (22%) outperforming deleteXYZ (30%) and deleteYall (44%). It is interesting to note that deleteXall (22%) outperforms deleteYall (44%) only when infection requires less than 15 minutes of contact time to spread.

Figure 8.7: Per edge strategy, we compute the average time taken (in hours) to infect a percentage of the network (5, 10, 20, 30, 40 %), where the average is computed over the number of nodes which achieved that level of potency. We plot the difference in average time taken, before and after edge deletion. A negative value means that, post edge-deletion, it took a shorter time (on average) to reach the level of potency. This may happen because in the edge-manipulated network, fewer nodes were able to reach that level of potency, driving down the average value. Legend labels are in decreasing order based on 5% potency.

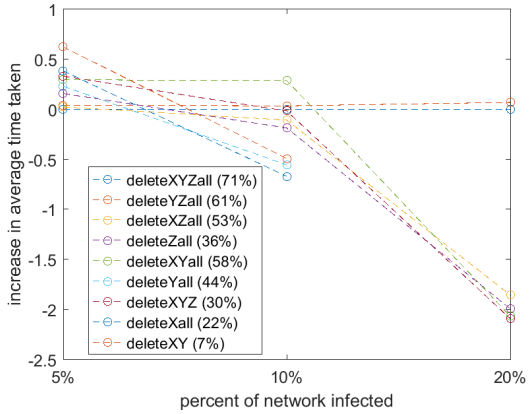
(a) 10 min



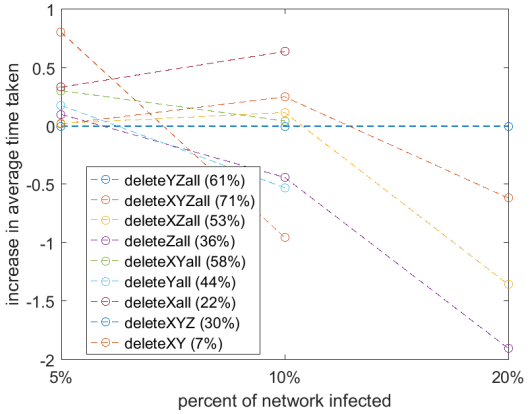
(b) 15 min



(c) 20 min



(d) 25 min



8.5 Conclusion

Targeted edge removal based on temporal centrality, while having a strong effect on structural properties of the network (as measured by dynamic total communicability DTC), has minimal effect on epidemic outcome. Our results show that edge manipulation reduces the number of potent nodes, but has on average a minimal effect on slowing down the epidemic spread. We conclude that temporal dynamics, while adding a layer of complexity to the analysis, also increases the robustness of the network with respect to epidemic spread. However, different conclusions may be reached if a different epidemic model is used.

On a structural level, our results show that nodes which are both high broadcasters and high receivers account for much of the interactions between the top-ranked nodes, and it is perhaps worthwhile to study how a combination of receive and broadcast scores/ranks can further enhance the analysis. It may also be instructive to compare the epidemic effect across the groups X, Y and Z.

Chapter 9

Conclusion

The overarching motivation for this work was to quantify the relationship between network-based centrality and contagion processes on a network. We were fortunate to have the opportunity to test our hypotheses on an empirical contact network based on the interactions of people in an ED of a busy, urban hospital.

Dynamic communicability succinctly uses matrix functions to capture temporal walk-based information and can be computed with ease. This method correctly identified staff members (specifically, RN's) as being at the receiving ends of 'short' walks, which is feasible due to their active role in the ED. This in itself provides reassurance that the theoretical framework underpinning dynamic communicability is capable of capturing something meaningful and interpretable from real data. However, our results also show that the method is unable to distinguish between a large majority of nodes. We have shown that in the case of receive centrality, this is not a result of normalization effects, but rather, a consequence of the large number of time-steps under study. It may be that such a measure is more useful when there are fewer time-steps, and some modifications must be performed to normalize the distribution of scores when the number of time-steps becomes too large. It is also possible that the modifications proposed in [35], which take into account the time of origin of

walks, can help to mitigate this effect.

An interesting observation is that while the top receivers are RN's, nodes that are both strong broadcasters *and* receivers are staff members that are *not* RN's. In addition, these nodes form a tightly-knit community, interacting often over the entire shift, and in a small subset of locations within the ED. It was also observed that a large majority of interactions (more than 70%) are accounted for by the nodes identified as highly central by dynamic communicability. Furthermore, removing the connections *between* highly central nodes had a stronger effect on structural connectivity of the network, more so than removing the nodes from the network entirely. These results suggest that dynamic communicability does indeed capture something meaningful and has potential for useful applications.

Strategic removal of connections between highly central nodes had a marked effect on the structural connectivity of the network, but this did not translate to a reduction in epidemic effect. This result is interesting in its own right, and suggests that contagion effects do not always have analogous counterparts in underlying network structure. However, it is important to bear in mind that our findings are highly dependent on the epidemic model used – it will not be surprising if a different disease transmission model produced different results. According to [40], 'appropriate observation time frames and good discrimination among types of potentially infectious contacts are viral in order for network analyses to be a valuable epidemiological tool. Our findings nonetheless suggest that contagion is extremely robust when temporal contact information is taken into account.

We showed that staff (as initial sources of infection) tend to be associated with higher epidemic outcome as compared to patients. This provides added impetus to consider *both* broadcasting and receiving scores in relation to epidemic spread. It may well be that broadcast centrality alone may not be a strong enough indicator of virulence. Regardless, regression analyses showed that centrality has a positive effect on epidemic outcome. We see that broadcast scores can provide information not captured by static/aggregated measures.

Additionally, we also observe that broadcasting ability is highly correlated with worst-case epidemics, suggesting that such measures can be useful in mitigating worst-case scenarios.

An important empirical finding is that broadcasting ability is largely dependent on time of entry. Consequently, time of appearance in the ED can be used as a proxy to the computation of broadcasting scores. The strong replication of broadcasting ranks, using the worst-case epidemic measure NS-EPI = EPI/$\log(t_s - t_a)$, provides another interpretation of broadcasting ability, and again reveals the dependence of this measure on activation time (since the saturation time, t_s, is similar for the majority of nodes).

With respect to other epidemic measures, the relationship to broadcast ranks is not so clear. Our results are not inconsistent with previous findings, that centrality scores based on network structure alone, are not always able to capture all top-spreaders: the spreading power of less central nodes may be underestimated. In addition, with respect to predicting epidemic outcome, temporal centrality does not necessarily outperform less complex measures. Nonetheless, temporal centrality identifies a distinct set of top-spreaders than centrality based on the time-aggregated binarized contact matrix, so that taken together, the accuracy of capturing top-spreaders improves significantly. Our results also show that other temporal markers (such as duration observed) can be used in a simple predictive model to generate predictions that capture the trend of the observed data remarkably well.

Much can also be gleaned from a non-network perspective. We showed that staff-staff interactions dominated all other types of interactions. Our study revealed 'hotspots' of interaction activity within the ED, such as the registration/triage area, the ED waiting areas, as well as a staff break area and office area. Temporal dynamics of interaction patterns at these locations can inform disease-prevention strategies, as well as design plans to optimize the utility and functionality of the ED as a whole.

In conclusion, while our study has shown that network analysis can reveal interesting insights into disease dynamics, epidemic spread is a complicated process which cannot be

fully explained by network structure. We have shown that while there is indeed some relationship between network-based centrality and epidemic outcome, immediate applications to mitigating epidemic effect remain nebulous. Inclusion of temporal information in the infection process is necessary to mimic real-world effects, but this added layer of complexity also increases the robustness of the network with respect to epidemic spread. Regardless, much more can be done to fine-tune and improve our analysis. Other temporal centrality measures can be investigated, and we believe that a combination of different network-based centralities, alongside other node-specific observables (such as patient/staff characteristics, duration observed, etc), can altogether paint a more complete picture of how contagion spreads on a network.

Appendix A: Partial lists of node rankings

We display partial lists of node rankings associated with various values of α. The first row of Table A.1 indicates that node 1 was ranked fifth when $\alpha = \alpha_1$, fourth when $\alpha = \alpha_2$, and sixth when $\alpha = \alpha_3, \alpha_4$. We compare the Pearson correlation between these lists (Table 4.3c) and visualize them in Figure 4.4.

Table A.1: BC

node	0.25	0.5	0.75	0.85
1	5	4	6	6
2	46	46	45	45
3	57	59	59	59
4	85	86	84	80
5	71	72	72	72
6	63	63	63	63
7	56	58	58	58
8	19	20	20	20
9	82	82	82	81
10	83	83	85	82
11	75	75	75	75
12	104	93	92	90
13	101	94	93	91
14	61	61	61	61
15	62	62	62	62
⋮	⋮	⋮	⋮	⋮

Table A.2: RC

node	0.25	0.5	0.75	0.85
1	21	20	20	20
2	83	68	57	52
3	1	3	2	4
4	8	8	8	8
5	16	16	16	16
6	9	9	9	9
7	10	10	10	10
8	19	21	21	21
9	4	4	5	5
10	15	15	15	15
11	12	12	11	11
12	26	26	26	26
13	27	27	27	27
14	6	5	3	1
15	51	51	51	51
⋮	⋮	⋮	⋮	⋮

Appendix B: Partial lists of nodes in ranked order

We display partial lists of nodes in ranked order associated with various values of α. For example, the first row of Table B.1 indicate that node 40 is ranked highest as a broadcaster regardless of α. Nodes highlighted in red are staff. We compute Kendall correlation (Table 4.3) and intersection distance *isim* (Table 4.4) as quantitative ways to assess similarity between the lists.

Table B.1: BC

rank	0.25	0.5	0.75	0.85
1	40	40	40	40
2	41	41	41	41
3	84	84	84	84
4	57	1	56	56
5	1	57	49	49
6	95	53	1	1
7	53	95	57	57
8	83	56	53	53
9	29	49	51	51
10	56	51	90	90
11	49	90	95	34
12	30	29	29	95
13	90	62	62	62
14	51	30	34	29
15	27	34	105	36
⋮	⋮	⋮	⋮	⋮

Table B.2: RC

rank	0.25	0.5	0.75	0.85
1	3	21	31	14
2	26	31	3	21
3	31	3	14	31
4	9	9	26	3
5	21	14	9	9
6	14	26	21	26
7	27	27	27	27
8	4	4	4	4
9	6	6	6	6
10	7	7	7	7
11	43	43	11	11
12	11	11	16	16
13	16	16	43	43
14	25	25	25	25
15	10	10	10	10
⋮	⋮	⋮	⋮	⋮

Appendix C: Pseudo-code for stochastic infection model

Consider a dynamic network represented by a sequence of (10-second aggregated) adjacency matrices $A^{[t]}$ for $t = 1, \ldots, 4321$. Each matrix is $n \times n$, where n is the number of nodes. We aim to generate an output consisting of a $n \times 5$ array of the form:

nodeID	infected	timeOfInfection	sourceOfInfection	# sinks
⋮	⋮	⋮	⋮	⋮

where the entries in the 'infected' column are boolean, indicating whether or not the node is infected; for infected nodes, the source of infection is the node which passed the infection directly to it; a sink corresponding to node i is a node that was directly infected by node i.

For a fixed time-step t, we work only with the (current) matrix $A^{[t]}$ and the previous matrix $A^{[t-1]}$. We scan the matrices (along each row) to see if any changes in connections took place at time t. Since the matrices are symmetric, we look only at the upper triangular part. In the process, we also keep track of the following:

- $toBeInfected$ vector which identifies the nodes that become infected at time t. Note that the 'infected' status of these nodes are only updated at the end of the scan. This is because infected nodes can only start infecting other nodes at the next time step.
- $TimeOfContact$ matrix where the ij-th entry is the length of time that the link between node i and node j is active. If node j is infected at time t, at the end of the scan, the entire column $TimeOfContact[\cdot][j]$ is reset to zero. This is because we want to start counting the connection times (with other nodes) only *after* node j becomes infected, and not before.

Note that $toBeInfected$ will be over-written (reset to $false$ vector) at each time-step. $TimeOfContact$ will be continually updated and the corresponding output will not be stored.

For a fixed time t, consider the ij-th entry of $A^{[t-1]}$ and $A^{[t]}$.

- **Case I: the ij-th entry is 0 at both times.**
  Do nothing.
- **Case II: the ij-th entry is 1 at both times, or the ijth entry changed from 0 to 1**
  Update $TimeOfContact[i][j] \leftarrow TimeOfContact[i][j] + 1$.
- **Case III: the ij-th entry changed from 1 to 0.**
  This means that a connection between nodes i and j ended at time t. We do two things:

1. **Check if infection occurs between i and j:** Check the 'infected' status of both nodes at time $t-1$. If both nodes have the same status, do nothing. If only one node is infected, compute the probability of infection using the length of time of contact, given by $TimeOfContact[i][j]$, and randomize to determine if infection occurs. If infection occurs, say node i infects node j, update the following:
   - $toBeInfected[j] \leftarrow true$
   - $timeOfInfection[j] \leftarrow t$
   - $sourceOfInfection[j] \leftarrow i$
   - $\#\text{sinks}[i] \leftarrow \#\text{sinks}[i] + 1$
2. **Reset the connection time between i and j:** $TimeOfContact[i][j] \leftarrow 0$. We can do this because after randomization, we no longer need this information. Furthermore, any future contact between the two nodes will be treated as independent.

After scanning every entry in $A^{[t-1]}$ and $A^{[t]}$,

- update the 'infected' status of the nodes based on the vector $toBeInfected$;
- reset the connection times in $TimeOfContact$ for all nodes infected at time t to zero.

Move on to the next time step.

We insert a zero matrix at the end of the adjacency sequence, to ensure that at the end of the time sequence, all pairs of nodes are tested for possible infection.

Algorithm 1 Infection Model Simulation

```
// Initialize
currentMatrix ← firstMatrix
previousMatrix ← firstMatrix
TimeOfContact ← zeroMatrix
U ← zeroMatrix

t ← 1.

loop:

toBeInfected ← falseVector                    ▷ This must be reset at each time-step

// Begin scan
i ← 1.
for i <= n do
    j ← i + 1.
    for j <= n do
        if TimeOfContact[i][j] > 0 then                          ▷ Connection active.
            if i is infected XOR j is infected then              ▷ Only one of two is infected
                // check if infection spread: compute probability; randomize
                if infected(TimeOfContact[i][j], U[i][j]) = true then   ▷ Infection spread
                    // Suppose i infected j (other case is also considered)
                    toBeInfected[j] ← true
                    timeOfInfection[j] ← t
                    sourceOfInfection[j] ← i
                    #sinks[i] ← #sinks[i] + 1
        j ← j + 1.
        // End inner for (finish scanning row i)
    i ← i + 1.
    // End outer for (finish scanning entire matrix)

// Update infection status and reset connection time for infected nodes
k ← 1.
for k <= n do
    if toBeInfected[k] = true then
        infected[k] ← true
        timeOfContact[·][k] ← 0                          ▷ Reset column k to zero
```

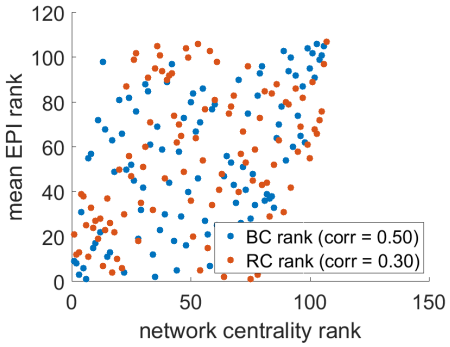
Algorithm 1 (continued)

```
// Update connection time for all nodes; see Case II & III
i ← 1.
for i <= n do
    j ← i + 1.
    for j <= n do
        if currentMatrix[i][j] = 1 then          ▷ Case II: There is a connection, increment time
            if TimeOfContact[i][j] = 0 then      ▷ Beginning of connection, assign random u
                U[i][j] ← randomDouble()
            TimeOfContact[i][j] ← TimeOfContact[i][j] + 1
        else                                     ▷ Case III: No connection, or connection ended at time t
            TimeOfContact[i][j] ← 0
        j ← j + 1.
    i ← i + 1.

t ← t + 1.

if currentMatrix = lastMatrix then
    goto end.
previousMatrix ← currentMatrix
currentMatrix ← nextMatrix
goto loop.
end
```

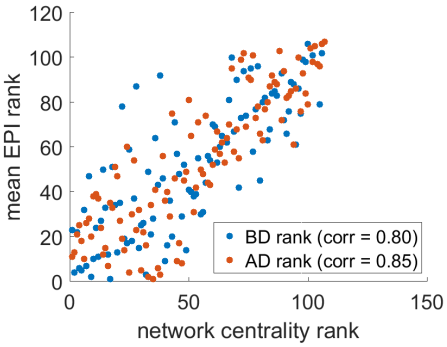
Appendix D: Mean EPI v centrality rankings

Figure D.1: Comparison of mean EPI rankings and network-based centrality rankings. Shifts 1-7.

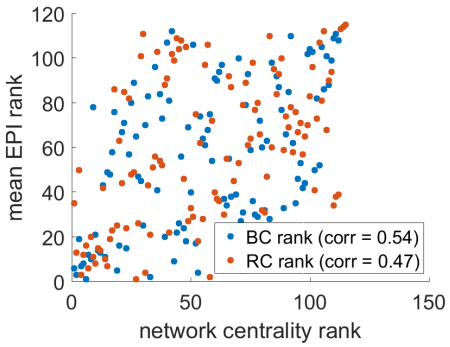
(a) Shift 1: Temporal



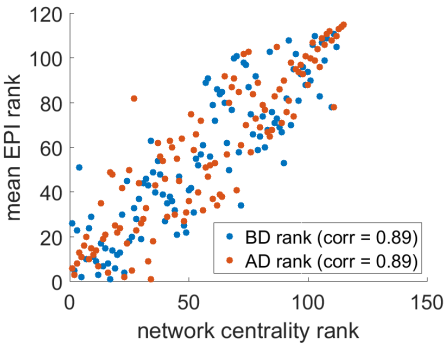
(b) Shift 1: Static



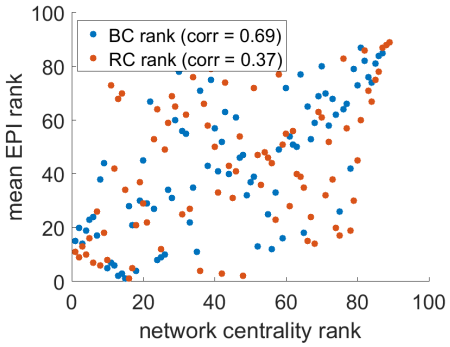
(c) Shift 2: Temporal



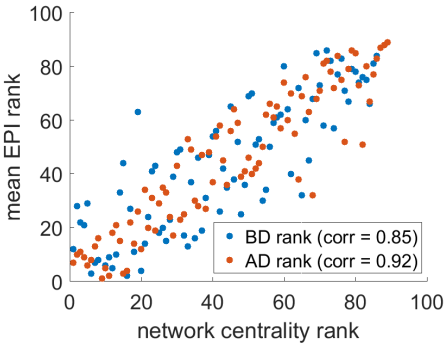
(d) Shift 2: Static



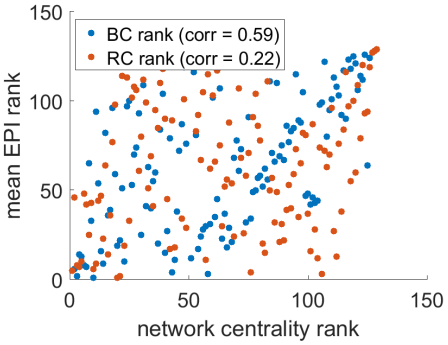
(e) Shift 3: Temporal



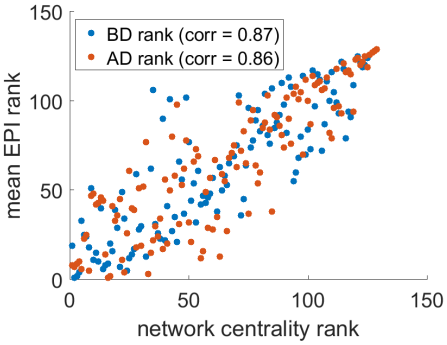
(f) Shift 3: Static



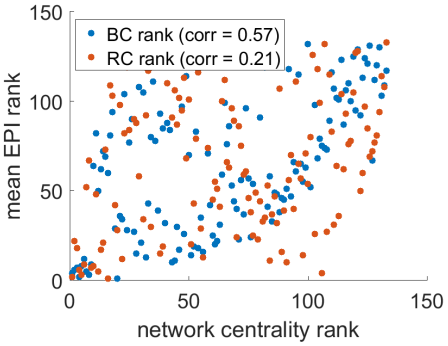
(a) Shift 4: Temporal



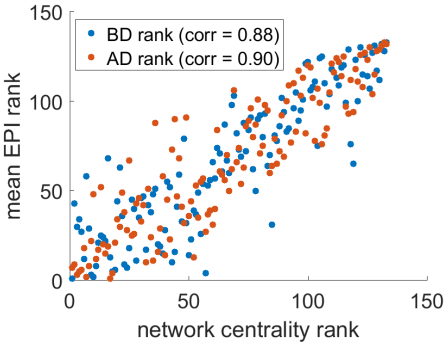
(b) Shift 4: Static



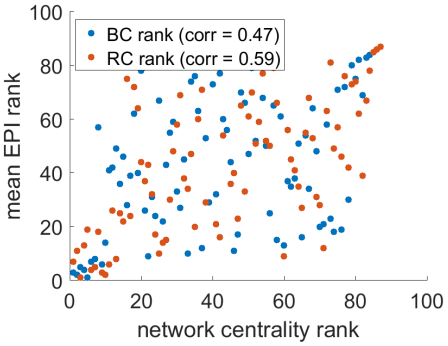
(c) Shift 5: Temporal



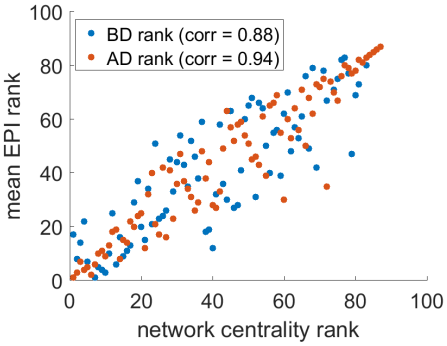
(d) Shift 5: Static



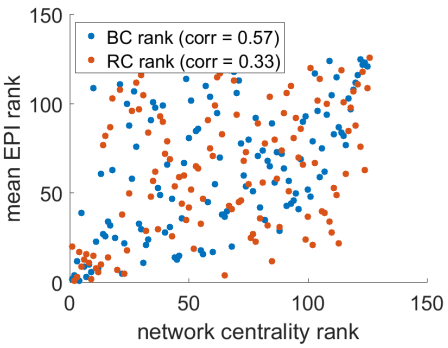
(e) Shift 6: Temporal



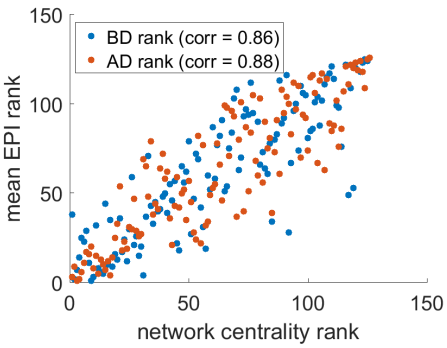
(f) Shift 6: Static



(g) Shift 7: Temporal



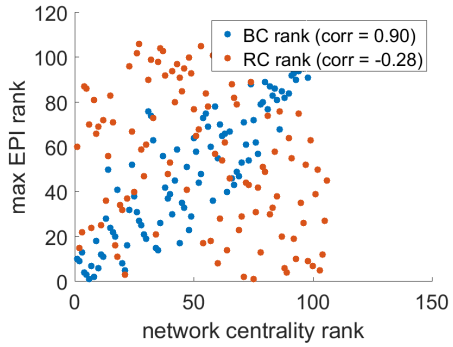
(h) Shift 7: Static



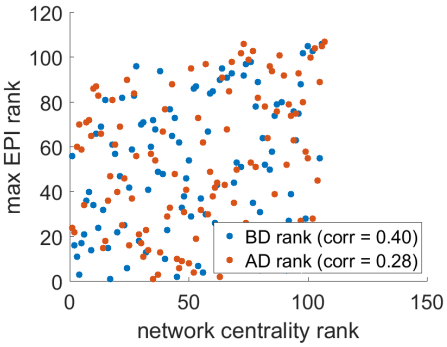
Appendix E: Max EPI v centrality rankings

Figure E.1: Comparison of max EPI rankings and network-based centrality rankings. Shifts 1-7.

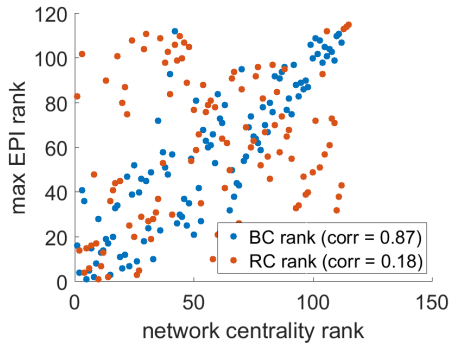
(a) Shift 1: Temporal



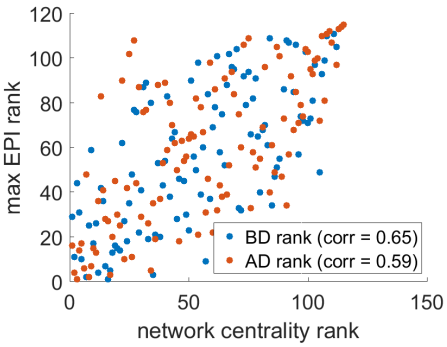
(b) Shift 1: Static



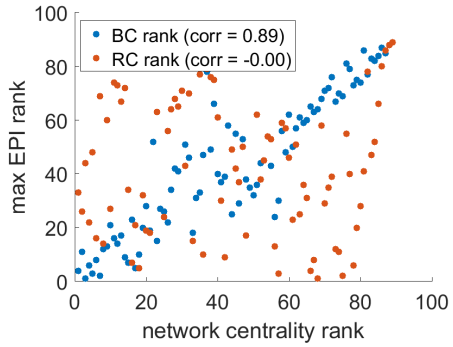
(c) Shift 2: Temporal



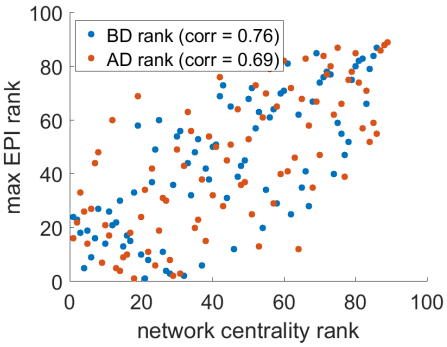
(d) Shift 2: Static



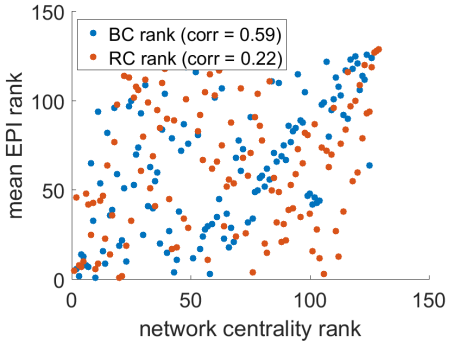
(e) Shift 3: Temporal



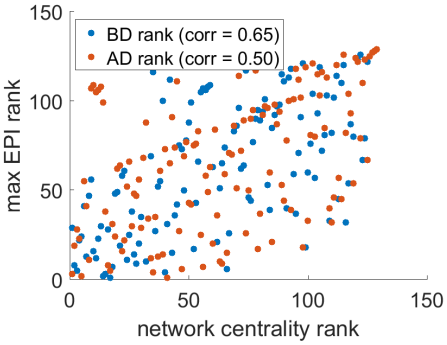
(f) Shift 3: Static



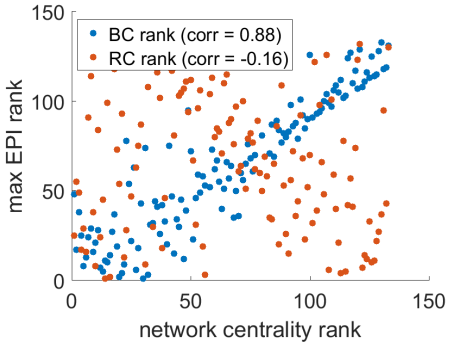
(a) Shift 4: Temporal



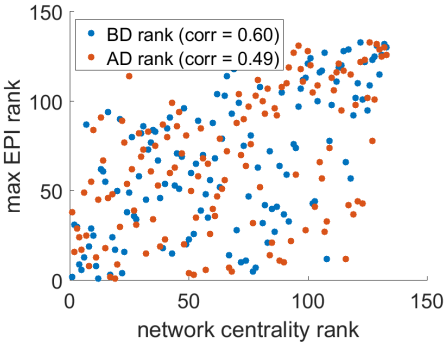
(b) Shift 4: Static



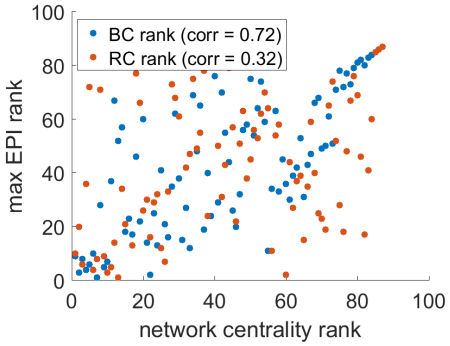
(c) Shift 5: Temporal



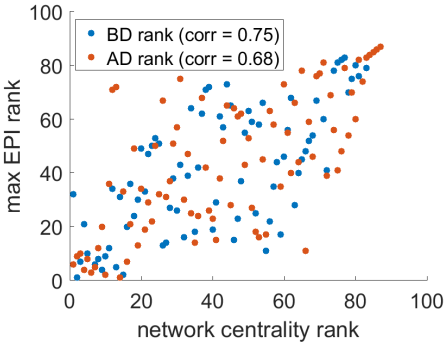
(d) Shift 5: Static



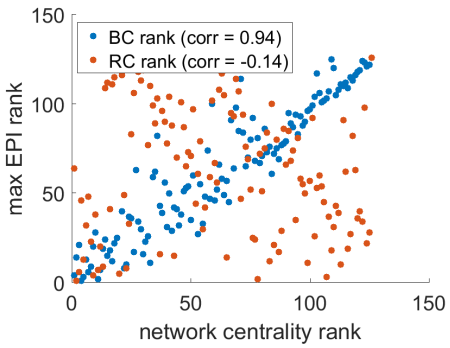
(e) Shift 6: Temporal



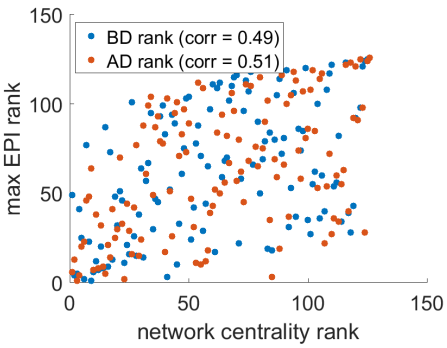
(f) Shift 6: Static



(g) Shift 7: Temporal



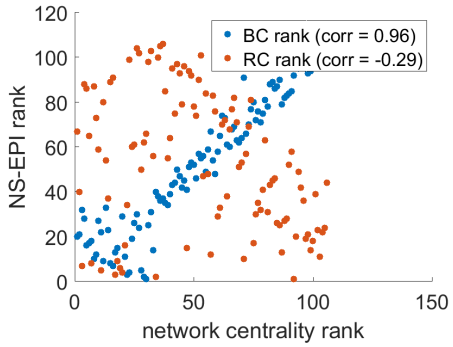
(h) Shift 7: Static



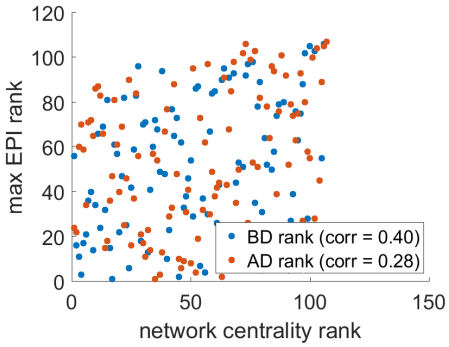
Appendix F: NS-EPI v centrality rankings

Figure F.1: Comparison of NS-EPI rankings and network-based centrality rankings. Shifts 1-7.

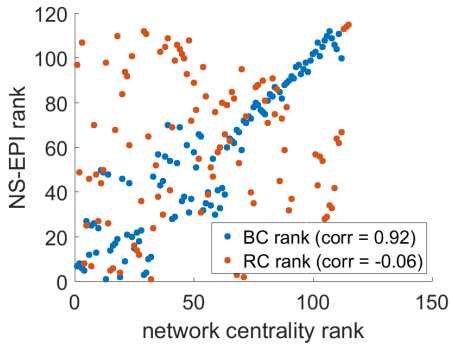
(a) Shift 1: Temporal



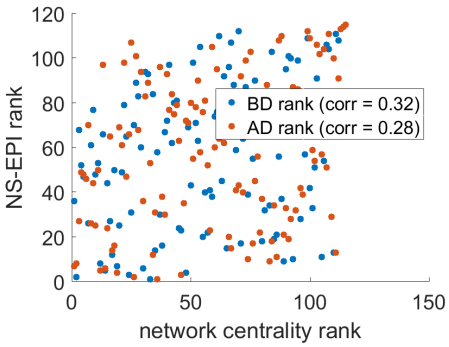
(b) Shift 1: Static



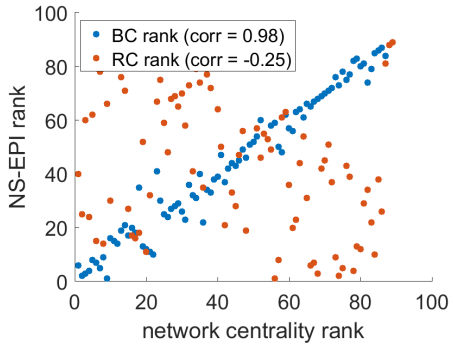
(c) Shift 2: Temporal



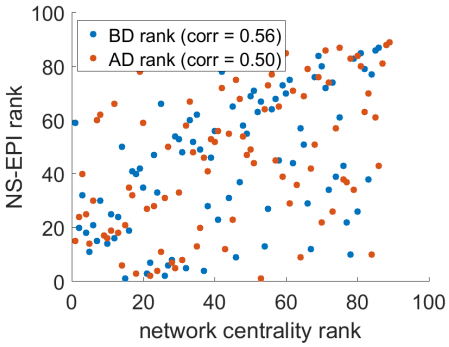
(d) Shift 2: Static



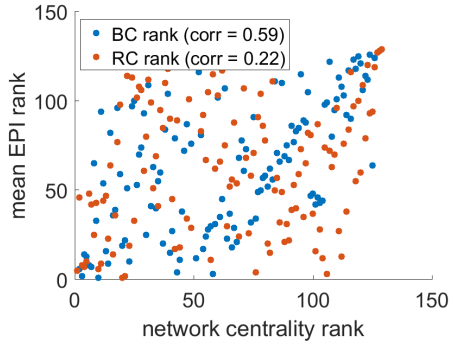
(e) Shift 3: Temporal



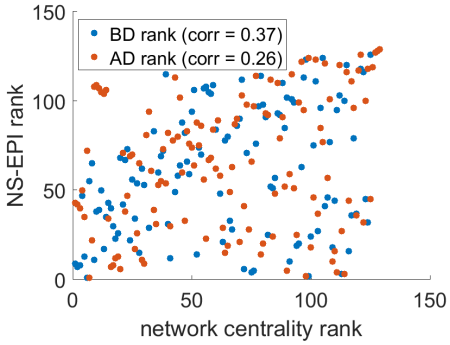
(f) Shift 3: Static



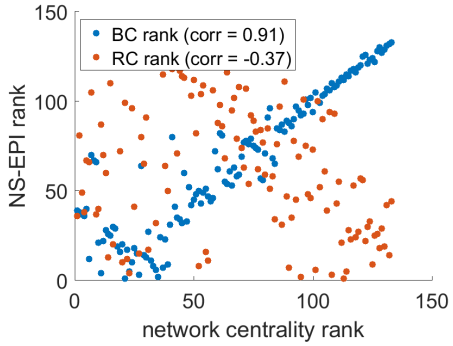
(a) Shift 4: Temporal



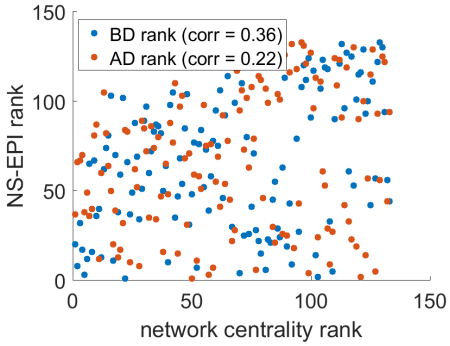
(b) Shift 4: Static



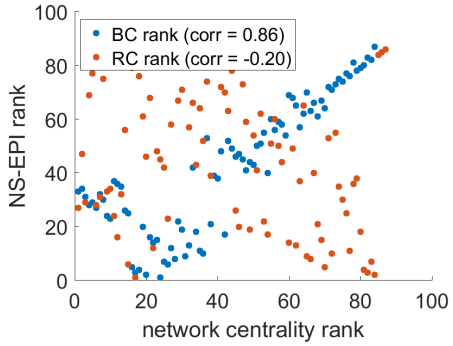
(c) Shift 5: Temporal



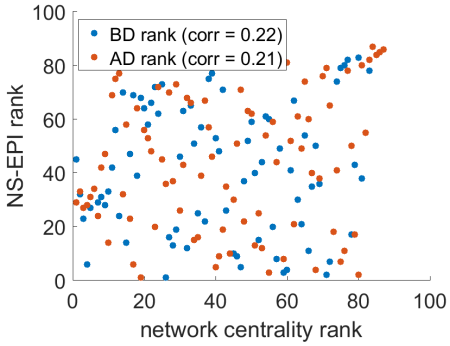
(d) Shift 5: Static



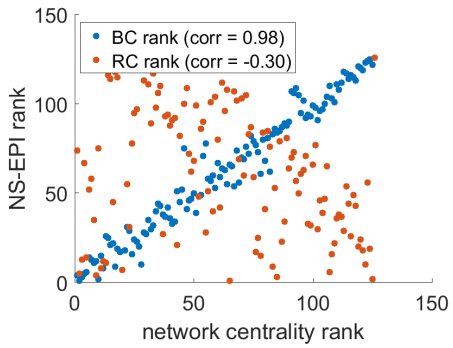
(e) Shift 6: Temporal



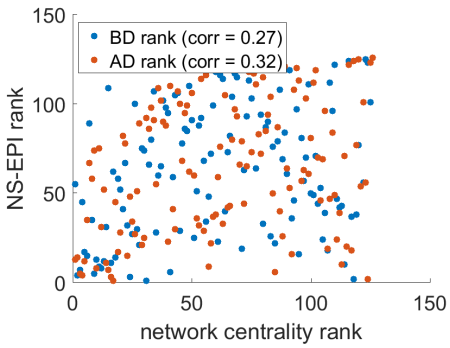
(f) Shift 6: Static



(g) Shift 7: Temporal



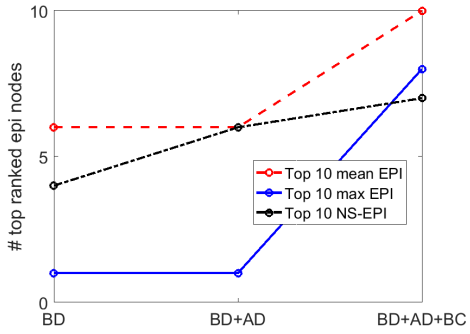
(h) Shift 7: Static



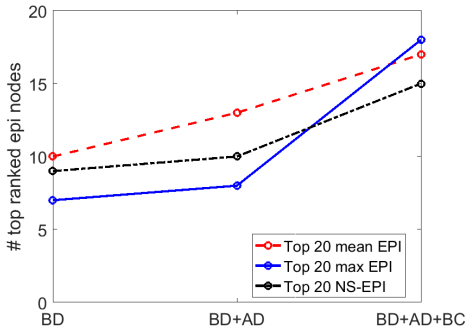
Appendix G: Added-value of BC

Figure G.1

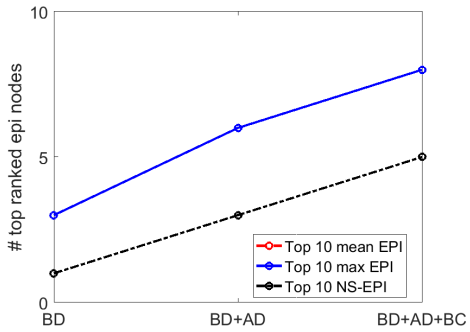
(a) Shift 1: Top 10



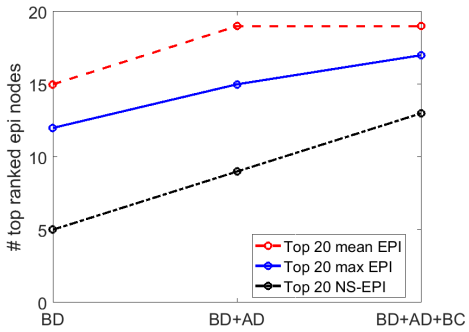
(b) Shift 1: Top 20



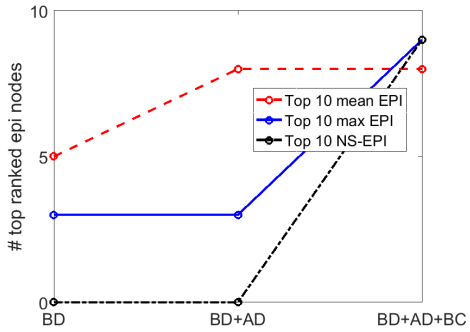
(c) Shift 2: Top 10



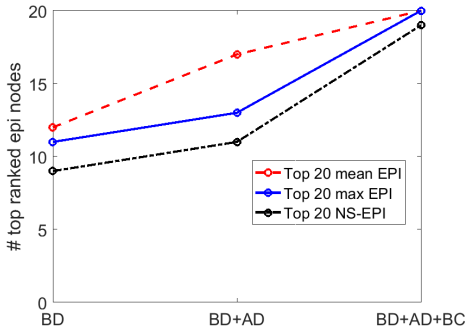
(d) Shift 2: Top 20



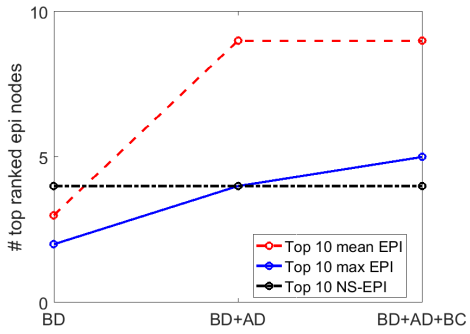
(e) Shift 3: Top 10



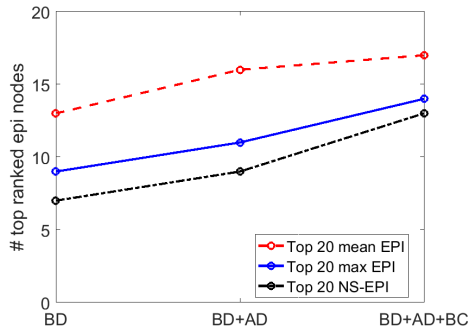
(f) Shift 3: Top 20



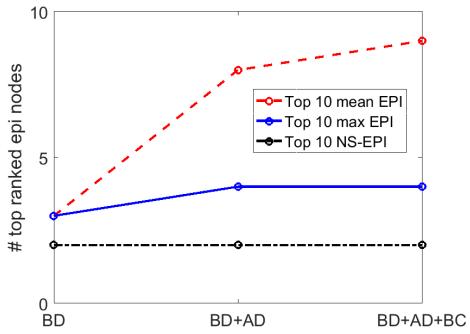
(a) Shift 4: Top 10



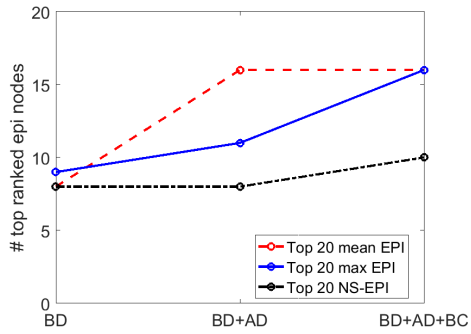
(b) Shift 4: Top 20



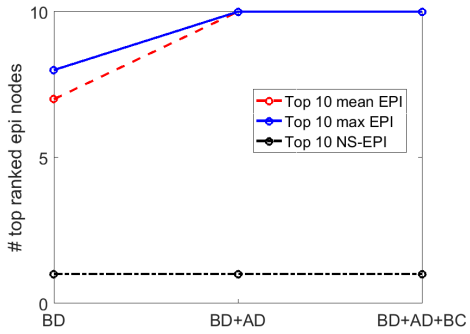
(c) Shift 5: Top 10



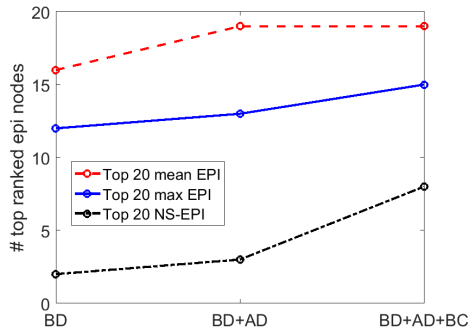
(d) Shift 5: Top 20



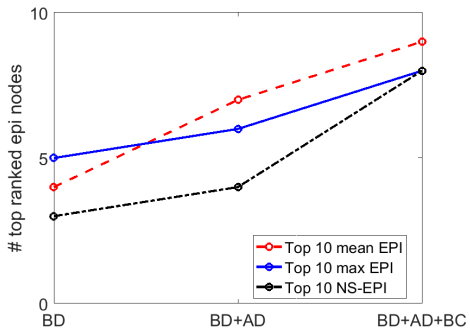
(e) Shift 6: Top 10



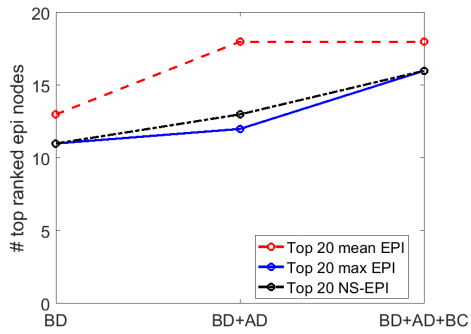
(f) Shift 6: Top 20



(g) Shift 7: Top 10



(h) Shift 7: Top 20



Appendix H: Predictions

We plot the predictions based on full models

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 S + \epsilon,$$

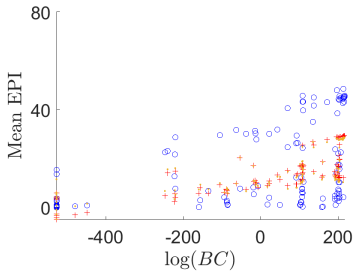
where Y is one of mean EPI, max EPI or NS-EPI, and X is one of log(BC), AD or BD. Coefficients are estimated based on Shift 1 data (training set). Predictions based on the null model

$$Y = \beta_0 + \beta_1 T + \beta_2 D + \beta_3 S + \epsilon$$

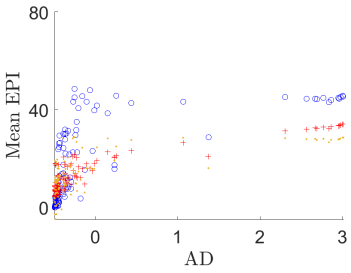
are included for comparison. Observed values are indicated by 'o', predicted values based on the null model are indicated by '·' and predicted values based on the full model are indicated by '+'.

Figure H.1: Predictions for Shift 2

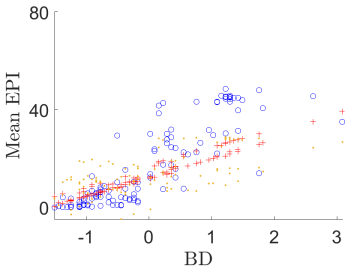
(a) mean EPI v log(BC)



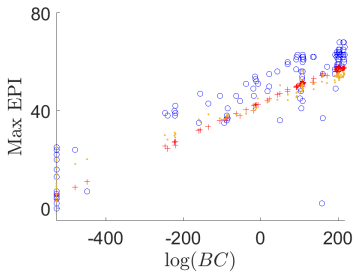
(b) mean EPI v AD



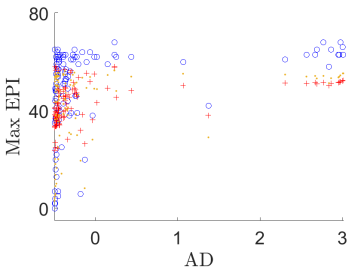
(c) mean EPI v BD



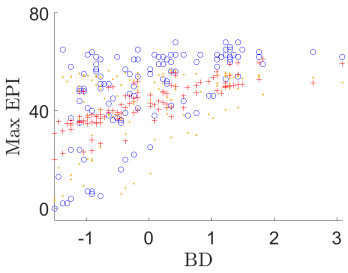
(d) max EPI v log(BC)



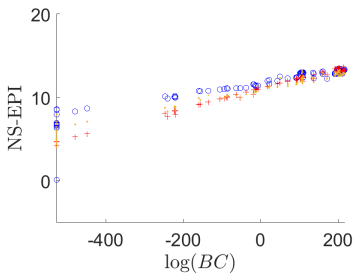
(e) max EPI v AD



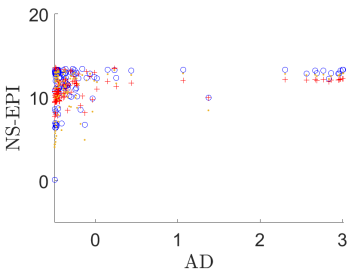
(f) max EPI v BD



(g) NS-EPI v log(BC)



(h) NS-EPI v AD



(i) NS-EPI v BD

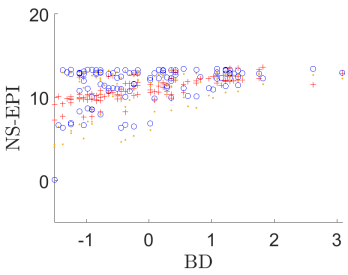
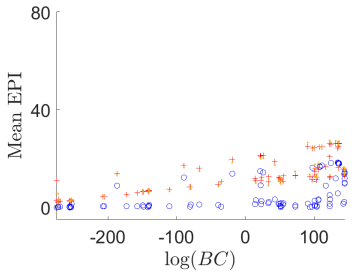
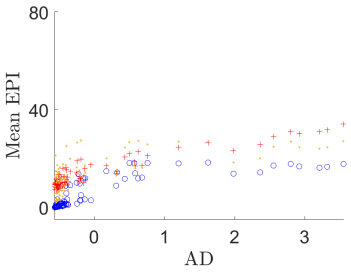


Figure H.2: Predictions for Shift 3

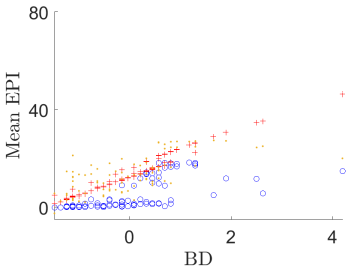
(a) mean EPI v log(BC)



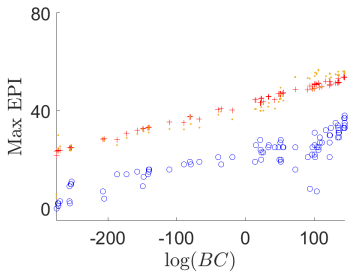
(b) mean EPI v AD



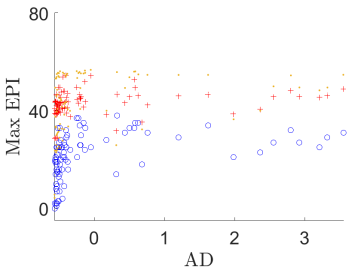
(c) mean EPI v BD



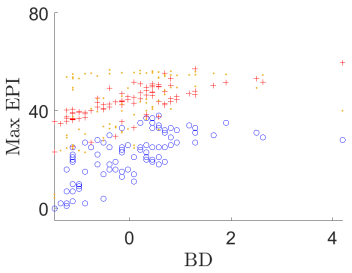
(d) max EPI v log(BC)



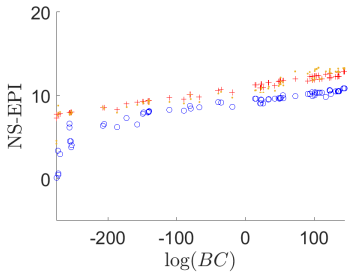
(e) max EPI v AD



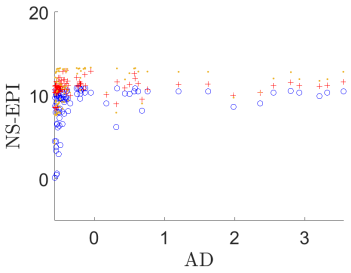
(f) max EPI v BD



(g) NS-EPI v log(BC)



(h) NS-EPI v AD



(i) NS-EPI v BD

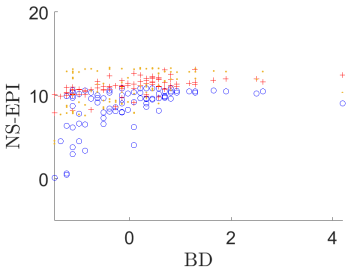


Figure H.3: Predictions for Shift 4

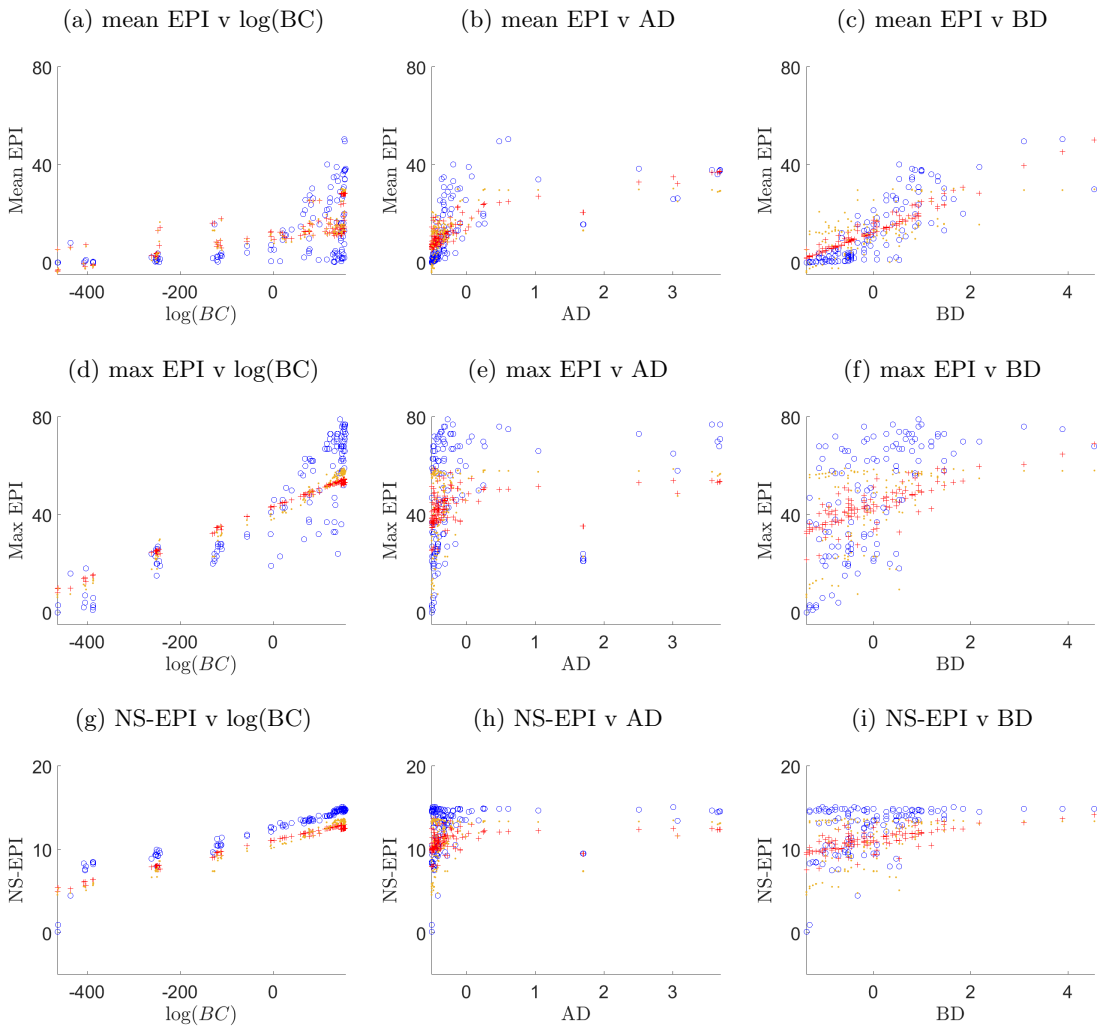


Figure H.4: Predictions for Shift 5

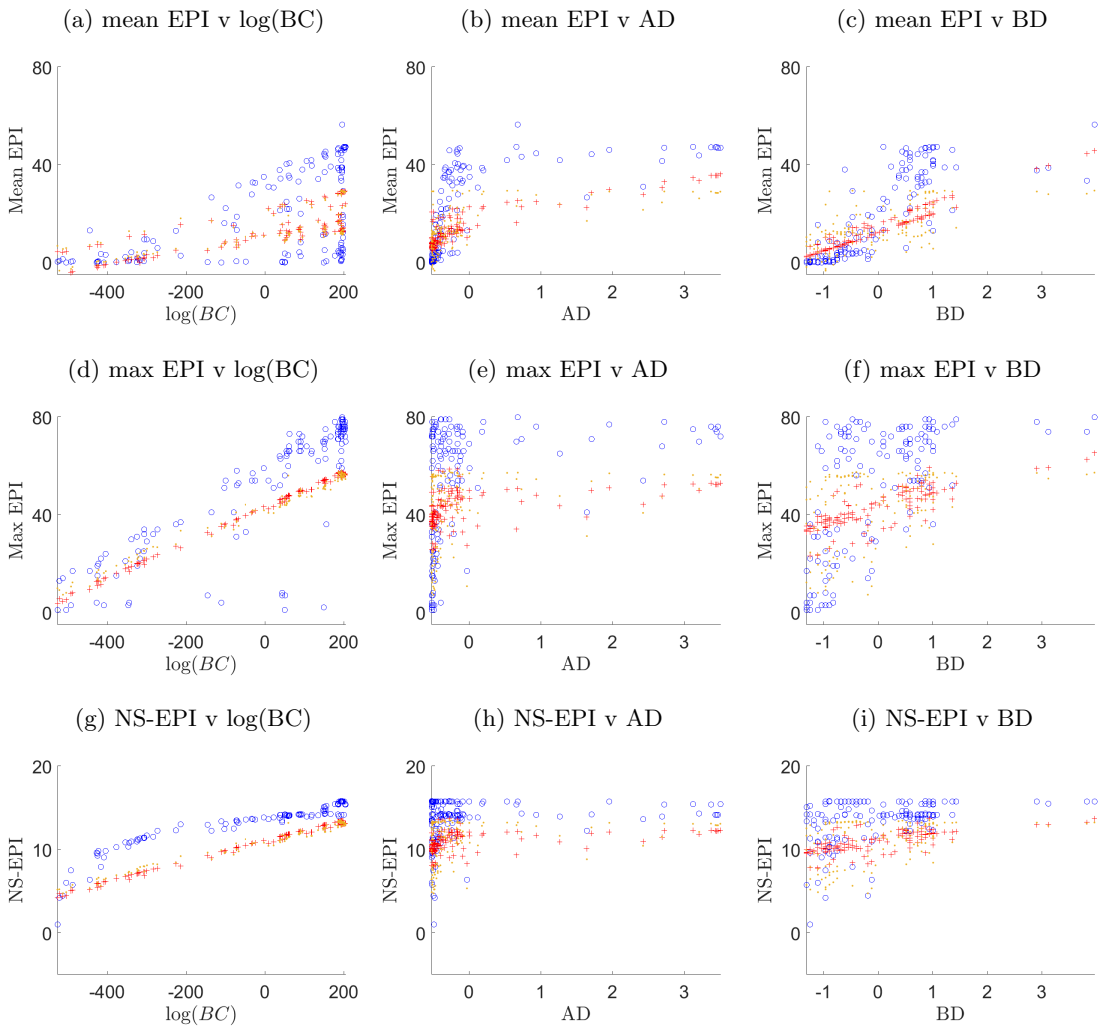
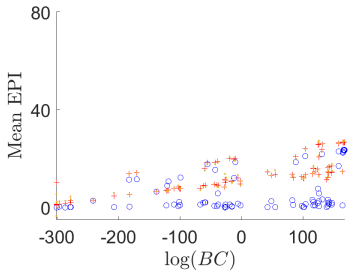
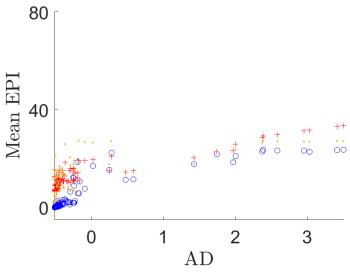


Figure H.5: Predictions for Shift 6

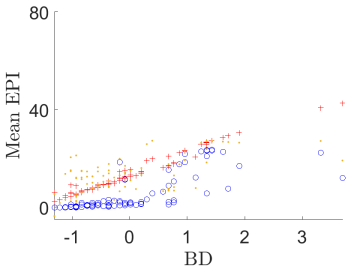
(a) mean EPI v log(BC)



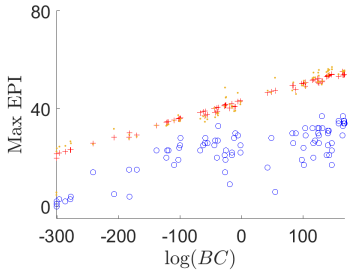
(b) mean EPI v AD



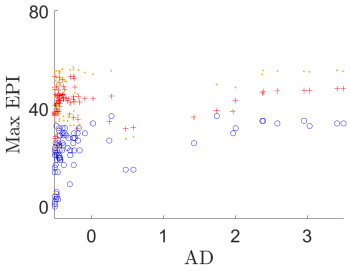
(c) mean EPI v BD



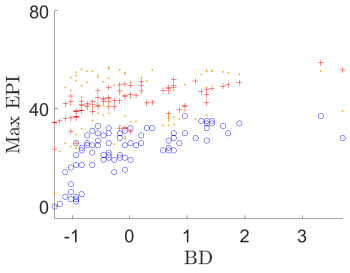
(d) max EPI v log(BC)



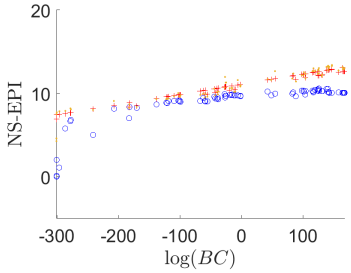
(e) max EPI v AD



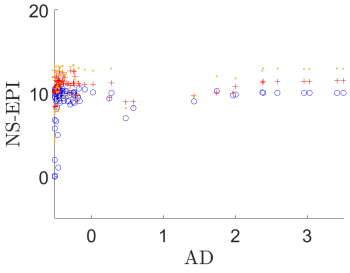
(f) max EPI v BD



(g) NS-EPI v log(BC)



(h) NS-EPI v AD



(i) NS-EPI v BD

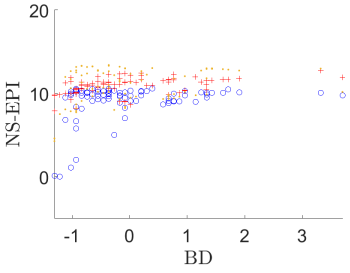
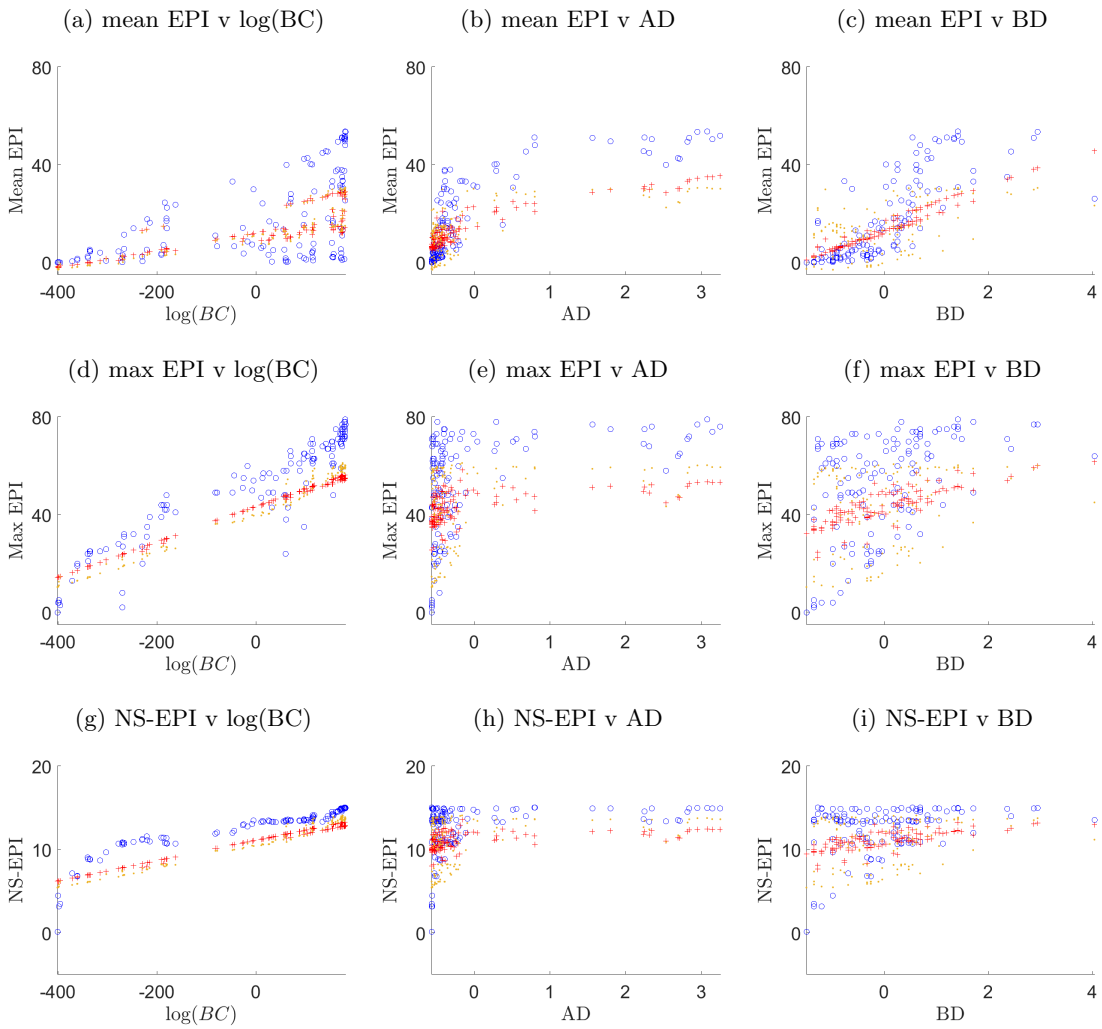


Figure H.6: Predictions for Shift 7



Bibliography

[1] Weihua An. Multilevel meta network analysis with application to studying network dynamics of network interventions. *Social Networks*, 43:48 – 56, 2015.

[2] L Ancel Meyers, M. E. J. Newman, M Martin, and S Schrag. Applying network theory to epidemics: Control measures for Mycoplasma pneumoniae outbreaks. *Emerging Infectious Diseases*, 9(2):204–10, 2003.

[3] Roy M. Anderson and Robert M. May. *Infectious Diseases of Humans, Dynamics and Control*. Oxford University Press, 1992.

[4] Francesca Arrigo and Michele Benzi. Edge modification criteria for enhancing the communicability of digraphs. *SIAM Journal on Matrix Analysis and Applications*, 37(1):443–468, 2016.

[5] Francesca Arrigo and Michele Benzi. Updating and downdating techniques for optimizing network communicability. *SIAM Journal on Scientific Computing*, 38(1):B25–B49, 2016.

[6] Guilherme Ferraz de Arruda, André Luiz Barbieri, Pablo Martín Rodríguez, Francisco A. Rodrigues, Yamir Moreno, and Luciano da Fontoura Costa. Role of centrality for the identification of influential spreaders in complex networks. *Phys. Rev. E*, 90(3):032812, Sep 2014.

[7] Norman T. J. Bailey. *Mathematical Theory of Epidemics*. Arnold, 1957.

[8] S Bansal, J Read, B Pourbohloul, and LA Meyers. The dynamic nature of contact networks in infectious disease epidemiology. *Journal of Biol Dyn*, 4(5):478–89, Sep 2010.

[9] Shweta Bansal, Bryan T Grenfell, and Lauren Ancel Meyers. When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of The Royal Society Interface*, 4(16):879–891, 2007.

[10] Shweta Bansal and Lauren Ancel Meyers. The impact of past epidemics on future disease dynamics. *Journal of Theoretical Biology*, 309:176 – 184, 2012.

[11] Shweta Bansal, Babak Pourbohloul, Nathaniel Hupert, Bryan Grenfell, and Lauren Ancel Meyers. The shifting demographic landscape of pandemic influenza. *PLoS ONE*, 5(2):1–8, 02 2010.

[12] Shweta Bansal, Babak Pourbohloul, and Lauren Ancel Meyers. A comparative analysis of influenza vaccination programs. *PLoS Med*, 3(10):1–10, 10 2006.

[13] CT Bauch. A versatile ode approximation to a network model for the spread of sexually transmitted diseases. *Journal of Mathematical Biology*, 45(5):375 – 395, Nov 2002.

[14] Frank Bauer and Joseph T. Lizier. Identifying influential spreaders and efficiently estimating infection numbers in epidemic models: A walk counting approach. *EPL (Europhysics Letters)*, 99(6):68007, 2012.

[15] Michele Benzi and Christine Klymko. Total communicability as a centrality measure. *Journal of Complex Networks*, 1(2):124–149, 2013.

[16] Michele Benzi and Christine Klymko. On the limiting behavior of parameter-dependent network centrality measures. *SIAM Journal on Matrix Analysis and Applications*, 36(2):686–706, 2015.

[17] Tanya Y Berger-Wolf. Graph theoretic measures for identifying effective blockers of spreading processes in dynamic networks. In *Proceedings of the MLG-ICML Workshop on Machine Learning on Graphs*, July 2008.

[18] Stephen P. Borgatti and Martin G. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484, 10 2006.

[19] Ciro Cattuto, Wouter Van der Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, 5(7):e11596, 2010.

[20] Davide Cellai and Ginestra Bianconi. Multiplex networks with heterogeneous activities of the nodes. *Phys. Rev. E*, 93:032302, Mar 2016.

[21] Ewan R. Colman and Nathaniel Charlton. Separating temporal and topological effects in walk-based network centrality. e-print arXiv:1601.00571v2, Jan 2016.

[22] Dragos Cvetković, Peter Rowlinson, and Slobodan Simić. *Eigenspaces of Graphs*. Cambridge University Press, 1997.

[23] Jos Antonio de la Pea, Ivan Gutman, and Juan Rada. Estimating the Estrada Index. *Linear Algebra and its Applications*, 427(1):70–76, 2007.

[24] Ken T. D. Eames and Matt J. Keeling. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proceedings of the National Academy of Sciences*, 99(20):13330–13335, 2002.

[25] P. Elliot, J. Wakefield, Nicola Best, and David Briggs. *Spatial Epidemiology: Methods and Applications*. Oxford University Press, 2001.

[26] Ernesto Estrada. *The Structure of Complex Networks, Theory and Applications*. Oxford University Press, 2012.

[27] Ernesto Estrada. Communicability in temporal networks. *Phys. Rev. E*, 88(4):042811, Oct 2013.

[28] Ernesto Estrada and Juan A. Rodríguez-Velázquez. Subgraph centrality in complex networks. *Phys. Rev. E*, 71:056103, May 2005.

[29] Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.

[30] H. Fefferman and K.L. Ng. How disease models in static networks can fail to approximate disease in dynamic networks. *Phys. Rev. E*, 76(3):031919, 2007.

[31] M. Carolyn Gates and Mark E.J. Woolhouse. Controlling infectious disease through the targeted manipulation of contact network structure. *Epidemics*, 12:11 – 19, 2015. Papers arising from Epidemics 4.

[32] A. V. Goltsev, S. N. Dorogovtsev, J. G. Oliveira, and J. F. F. Mendes. Localization and spreading of diseases in complex networks. *Phys. Rev. Lett.*, 109(12):128702, 2012.

[33] Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453:779 – 782, June 2008.

[34] Danica Vukadinovic Greetham, Zhivko Stoyanov, and Peter Grindrod. On the radius of centrality in evolving communication networks. *Journal of Combinatorial Optimization*, 28(3):540–560, February 2014.

[35] Peter Grindrod and Desmond J. Higham. A matrix iteration for dynamic network summaries. *SIAM Review*, 55(1):118–128, 2013.

[36] Peter Grindrod and Desmond J. Higham. A dynamical systems view of network centrality. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 470(2165), 2014.

[37] Peter Grindrod, Mark C. Parsons, Desmond J. Higham, and Ernesto Estrada. Communicability across evolving networks. *Phys. Rev. E*, 83(4):046120, Apr 2011.

[38] W. H. Hamer. *Epidemic Disease in England: The Evidence of Variability and of Persistency of Type*. Bedford Press, 1906.

[39] Petter Holme. Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(234), September 2015.

[40] Rowland R Kao, Darren M Green, Jethro Johnson, and Istvan Z Kiss. Disease dynamics over very different time-scales: foot-and-mouth disease and scrapie on the network of livestock movements in the uk. *Journal of the Royal Society Interface*, 4(16):907–916, 2007.

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

[42] M.J. Keeling and K.T.D. Eames. Networks and epidemic models. *J. R. Soc. Interface*, 2(4):295–307, Sep 2005.

[43] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A*, 115(772), 1927.

[44] Hyoungshick Kim and Ross Anderson. Temporal node centrality in complex networks. *Phys. Rev. E*, 85(2):026107, Feb 2012.

[45] Maksim Kitsak, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernan A. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, Nov 2010.

[46] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 2014.

[47] Konstantin Klemm, M. Ángeles Serrano, Vctor M. Eguluz, and Maxi San Miguel. A measure of individual role in f. *Scientific Reports*, 2:292, Feb 2012.

[48] AB Lawson. *Statistical Methods in Spatial Epidemiology*. Wiley, 2006.

[49] Glenn Lawyer. Understanding the influence of all nodes in a network. *Scientific Reports*, 5:8665, Mar 2015.

[50] Douglas W. Lowery-North, Vicki Stover Hertzberg, Lisa Elon, George Cotsonis, Sarah A. Hilton, Christopher F. Vaughns II, Eric Hill, Alok Shrestha, Alexandria Jo, and Nathan Adams. Measuring social contacts in the emergency department. *PLoS ONE*, 8(8):e70854, 2013.

[51] Alexander V. Mantzaris, Danielle S. Bassett, Nicholas F. Wymbs, Ernesto Estrada, Mason A. Porter, Peter J. Mucha, Scott T. Grafton, and Desmond J. Higham. Dynamic network centrality summarizes learning in the human brain. *Journal of Complex Networks*, 1(1):83–92, 2013.

[52] Alexander V. Mantzaris and Desmond J. Higham. A model for dynamic communicators. *European Journal of Applied Mathematics*, 23(6):659–668, Dec 2012.

[53] Alexander V. Mantzaris and Desmond J. Higham. Dynamic communicability predicts infectiousness. In Petter Holme and Jari Saramki, editors, *Temporal Networks*, Understanding Complex Systems, pages 283–294. Springer Berlin Heidelberg, 2013.

[54] Lauren Ancel Meyers, Babak Pourbohloul, M.E.J. Newman, Danuta M. Skowronski, and Robert C. Brunham. Network theory and SARS: predicting outbreak diversity. *Journal of Theoretical Biology*, 232(1):71 – 81, 2005.

[55] MR Moser et al. An outbreak of influenza aboard a commercial airliner. *American Journal of Epidemiology*, 110(1):1–6, Jul 1979.

[56] Joël Mossong et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 5(3):e74, 2008.

[57] M. E. J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):016128, July 2002.

[58] M. E. J. Newman. Properties of highly clustered networks. *Phys. Rev. E*, 68:026121, Aug 2003.

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

[60] Shogo Osawa and Tsuyoshi Murata. Selecting seed nodes for influence maximization in dynamic networks. In Giuseppe Mangioni, Stephen Miles Uzzo, Filippo Simini, and Dashun Wang, editors, *Complex Networks VI: Proceedings of the 6th Workshop on Complex Networks*, pages 91–98. Springer, 2015.

[61] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3203, Apr 2001.

[62] Gail E Potter, Mark S Handcock, Ira M Longini, and M Elizabeth Halloran. Estimating within-school contact networks to understand influenza transmission. *The Annals of Applied Statistics*, 6(1):1–26, March 2012.

[63] JM Read, KT Eames, and Edmunds WJ. Dynamic social networks and the implications for the spread of infectious disease. *J R Soc Interface*, 5(26):1001–1007, 2008.

[64] Sebastián Risau-Gusman. Influence of network dynamics on the spread of sexually transmitted diseases. *Journal of the Royal Society Interface*, 9(71):1363 – 1372, 2011.

[65] Luis E. C. Rocha, Frederik Liljeros, and Petter Holme. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Computational Biology*, 7(3):e1001109, 2011.

[66] Pejman Rohani, Xue Zhong, and Aaron A King. Contact network structure explains the changing epidemiology of pertussis. *Science*, 330(6006):982–985, 2010.

[67] JN Rosenquist, JH Fowler, and NA Christakis. Social network determinants of depression. *Molecular Psychiatry*, 16:273–281, 2011.

[68] Ronald Ross. Some quantitative studies in epidemiology. *Nature*, 87(2188):466 – 467, 1911.

[69] Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W. Feldman, and James H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025, 2010.

[70] Holly B. Shakya, Nicholas A. Christakis, and James H. Fowler. Social network predictors of latrine ownership. *Social Science & Medicine, Special Issue: Social Networks, Health and Mental Health*, 125:129–138, 2015.

[71] Renato Aparecido Pimentel da Silva, Matheus Palhares Viana, and Luciano da Fontoura Costa. Predicting epidemic outbreak from individual features of the spreaders. *Journal of Statistical Mechanics: Theory and Experiment*, P07005, 2012.

[72] Timo Smieszek. A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread. *Theor Biol Med Model*, 6(1):25–35, 2009.

[73] Timo Smieszek, Lena Fiebig, and Roland W. Scholz. Models of epidemics: when contact repetition and clustering should be included. *Theoretical Biology and Medical Modelling*, 6(1):1–15, 2009.

[74] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Corinne Régis, Jean-François Pinton, Nagham Khanafer, Wouter Van den Broeck, and Philippe Vanhems. Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BioMed Central Medicine*, 9(1):87–102, 2011.

[75] John Tang, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Vincenzo Nicosia. Analysing information flows and key mediators through temporal centrality metrics. In *Proceedings of the 3rd Workshop on Social Network Systems*, SNS '10, pages 1–6, New York, NY, USA, 2010. ACM.

[76] Dane Taylor, Sean A. Myers, Aaron Clauset, Mason A. Porter, and Peter J. Mucha. Eigenvector-based centrality measures for temporal networks. Under review at Multiscale Modeling and Simulation: A SIAM Interdisciplinary Journal (2015).

[77] Damon J. A. Toth, Molly Leecaster, Warren B. P. Pettey, Adi V. Gundlapalli, Hongjiang Gao, Jeanette J. Rainey, Amra Uzicanin, and Matthew H. Samore. The role of heterogeneity in contact timing and duration in network models of influenza spread in schools. *Journal of The Royal Society Interface*, 12(108):20150279, 2015.

[78] B.A.N. Travenolo and L. da F. Costa. Accessibility in complex networks. *Physics Letters A*, 373(1):89–95, 2008.

[79] Mile Šikić, Alen Lančić, Nino Antulov-Fantulin, and Hrvoje Štefančić. Epidemic centrality is there an underestimated epidemic impact of network peripheral nodes? *The European Physical Journal B*, 86(10):440, 2013.

[80] Matheus P. Viana, João L. B. Batista, and Luciano da F. Costa. Effective number of accessed nodes in complex networks. *Phys. Rev. E*, 85(3):036105, Mar 2012.

[81] Erik Volz and Lauren Ancel Meyers. Susceptible-infected-recovered epidemics in dynamic contact networks. *Proc. R. Soc. B*, 274(1628):2925–2933, 2007.

[82] Erik M. Volz, Joel C. Miller, Alison Galvani, and Lauren Ancel Meyers. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. *PLoS Comput Biol*, 7(6):1–13, 06 2011.

[83] Thomas P. Weber and Nikolaos I. Stilianakis. Inactivation of influenza a viruses in the environment and modes of transmission: A critical review. *Journal of Infection*, 57(5):361 – 373, 2008.