The Dissertation Committee for Lauren Marie Gardner Certifies that this is the approved version of the following dissertation:

NETWORK BASED PREDICTION MODELS FOR COUPLED TRANSPORTATION-EPIDEMIOLOGICAL SYSTEMS

Committee:

S. Travis Waller, Supervisor

Sahotra Sarkar

C. Michael Walton

Leon Lasdon

Zhanmin Zhang

Ivan Damnjanovic

NETWORK BASED PREDICTION MODELS FOR COUPLED TRANSPORTATION-EPIDEMIOLOGICAL SYSTEMS

by

Lauren Marie Gardner, B.S.Arch.E.; M.S.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2011

To my parents

Acknowledgements

I am indebted to countless people for their support throughout my graduate career, culminating in the completion of this dissertation. Firstly I want to extend my sincerest gratitude to my advisor, Dr. Waller, for his patience, confidence, and friendship over the last seven years, and for advising me well beyond issues specific to my research. I am greatly appreciative of Dr. Sarkar for sharing his knowledge of dengue, species distribution models, and not making me go collect triatomines. I would also like to thank the rest of my dissertation committee, Dr. Walton, Dr. Zhang, Dr. Damnjanovic and Dr. Lasdon for contributing their time, expertise and perspective towards this broad research endeavor.

The relationships I developed with individuals both within and outside of the department were highlights of my graduate school experience. I want to thank my friends for both putting up with and appreciating my inability to stop asking questions. In addition to his friendship, I could not be more grateful for David's time and contributions during our many coding and brainstorming sessions. Possibly just as integral to my graduation success were the countless coffee breaks taken with Roshan, Nezam and David, further enlightened by our inexhaustible set of conversation topics. In addition to the rest of the Waller research team, I am totally grateful for the advice and guidance I have continued to receive from some recent graduates, Jen, Avi, Nati and Steve. I would also like to thank Libbie, Lisa, and Vicki for their invaluable assistance on numerous occasions.

Lastly I thank my family for teaching me how to think, how to question, and providing me with the opportunity and encouragement to pursue my academic interests.

Lauren Marie Gardner

The University of Texas at Austin
May 2011

NETWORK BASED PREDICTION MODELS FOR COUPLED TRANSPORTATION-EPIDEMIOLOGICAL SYSTEMS

Publication No._____________

Lauren Marie Gardner, Ph.D.
The University of Texas at Austin, 2011

Supervisor: S. Travis Waller

The modern multimodal transportation system provides an extensive network for human mobility and commodity exchange around the globe. As a consequence these interactions are often accompanied by disease and other biological infectious agents. This dissertation highlights the versatility of network models in quantifying the combined impact transportation systems, ecological systems and social networks have on the epidemiological process. A set of predictive models intended to compliment the current mathematical and simulation based modeling tools are introduced. The main contribution is the incorporation of dynamic infection data, which is becoming increasingly available, but is not accounted for in previous epidemiological models. Three main problems are identified.

The objective of the first problem is to identify the path of infection (for a specific disease scenario) through a social contact network by invoking the use of network based optimization algorithms and individual infection reports. This problem parallels a novel and related problem in phylodynamics, which uses genetic sequencing data to reconstruct the most likely spatiotemporal path of infection.

The second problem is a macroscopic application of the methodology introduced in the first problem. The new objective is to identify links in a transportation network responsible for spreading infection into new regions (spanning from a single source) using regional level infection data (e.g. when the disease arrived at a new location). The new network structure is defined by nodes which represent regions (cites, states, countries) and links representing travel routes.

The third research problem is applicable to vector-borne diseases; those diseases which are transmitted to humans through the bite of an infected vector (i.e. mosquito), including dengue and malaria. The role of the vector in the infection process inherently alters the spreading process (compared to human contact diseases), which must be addressed in prediction models. The proposed objective is to quantify the risk posed by air travel in the global spread of these types of diseases.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION AND MOTIVATION

The continual expansion of the multi-modal global transportation network is constantly shrinking the distance between any two points on earth. Such a physically connected world is advantageous for a variety of reasons, opening doors for communication, education and commodity exchange with societies in almost any corner of the world. However, an extensive transportation network also poses new threats to the modern world; superseding natural geographic barriers previously responsible for containing infectious agents, and bridging (previously) isolated regions. As a result, both travelers and commodities are accompanied by a variety of infectious agents, disguised in the form of humans and animals (i.e. insects, bacteria, and parasites among others), dispersing new and old diseases around the globe. This research highlights the versatility of network models in quantifying the combined impact transportation systems, ecological systems and social networks have on the epidemiological process. In this research an epidemiological system refers to any system in which a disease spreading process can be formulated (i.e., social-contact networks, transportation systems).

The impact of transportation on spreading infection has been observed on countless occasions throughout history. Possibly the most notorious example is the Black Death (Bubonic Plague) that swept through Europe in the 14th century and killed an estimated 75 million people, or 30-60% of the European population. The plague is thought to have been brought into southern Europe via (infected fleas on rats on board) ships as part of a trade route (Hayes, 2005); the most likely origin of the disease was recently identified as China (Wade, 2010). A more recent example is the worldwide 1918 flu pandemic (called the Spanish Flu because Spanish King Alfonso XIII became gravely ill and was the highest-profile patient about whom there was coverage, hence the widest and most reliable news coverage came from Spain). This pandemic lasted from March 1918 to June 1920, even spreading to remote destinations such as the Arctic and Pacific islands. The death toll is estimated between 50 and 100 million; and 500 million (an estimated 1/3 of the world's population) were infected, ranking as one of the deadliest natural disasters in human history (CDC, 2009). While the source is still undefined, the increased travel of soldiers, sailors, and civilians aided by modern transportation systems is recognized as a significant factor in the worldwide occurrence of this flu. Additionally the close troop quarters and massive troop movements hastened the pandemic and probably both increased transmission and augmented mutation. These catastrophic examples of disease transmission across space and time are not the first or last; the threat of global pandemics will continue to increase concurrently with our emergent global transportation systems.

1.1 Challenges

Many researchers have sought methods to predict and prevent the spread of various types of infectious agents within and between communities. Regardless of the spreading agent (human, mosquito, water-borne parasite), this is an extremely complex problem for multiple reasons. The greatest challenge is posed by the inevitable interdisciplinary nature of the problem; drawing on applications from biology, sociology, mathematics, statistics, anthropology, psychology, policy and engineering. For example, a communicable disease which spreads via human-to-human contact will spread locally through a social network (defined by individual's daily activity patterns, which are further dependent on local transportation and behavior patterns); and will spread inter-regionally through a transportation network (which could be air, rail, shipping, or some combination thereof). The size of an outbreak is determined (in part) by the biological components of the disease and the opportunity for contagion; while the success of control measures is based (in part) on the policies implemented and human compliance thereafter.

In terms of modeling, additional complications arise due to the stochastic nature of the problem; specifically: 1) the stochastic nature of most diseases' infection processes (e.g. it is not known with certainty when and where most diseases are transmitted); 2) uncertainty in the structure of the (social) network (e.g., full contact information is not available to researchers, and many human interactions cannot be accurately predicted); 3) incomplete data sets (e.g., many infections are reported inaccurately or at all); and 4) various other uncertainties inherent in the daily world.

To further complicate the problem, infection processes occur in parallel across different network systems. For example a disease can spread locally via humans in a social network, regionally via humans (or other biological agents) in a transportation network, and perhaps geographically via infectious biological agents. The problem of integrating the various network structures in conjunction with the stochastic processes taking place on each is a research topic addressed in this work.

In order to accurately predict how diseases might evolve across space and time it is necessary to account for the various disciplinary facets, the integrated problem structure and the stochastic nature of the problem.

These challenges are summed up in Figure 1-1.

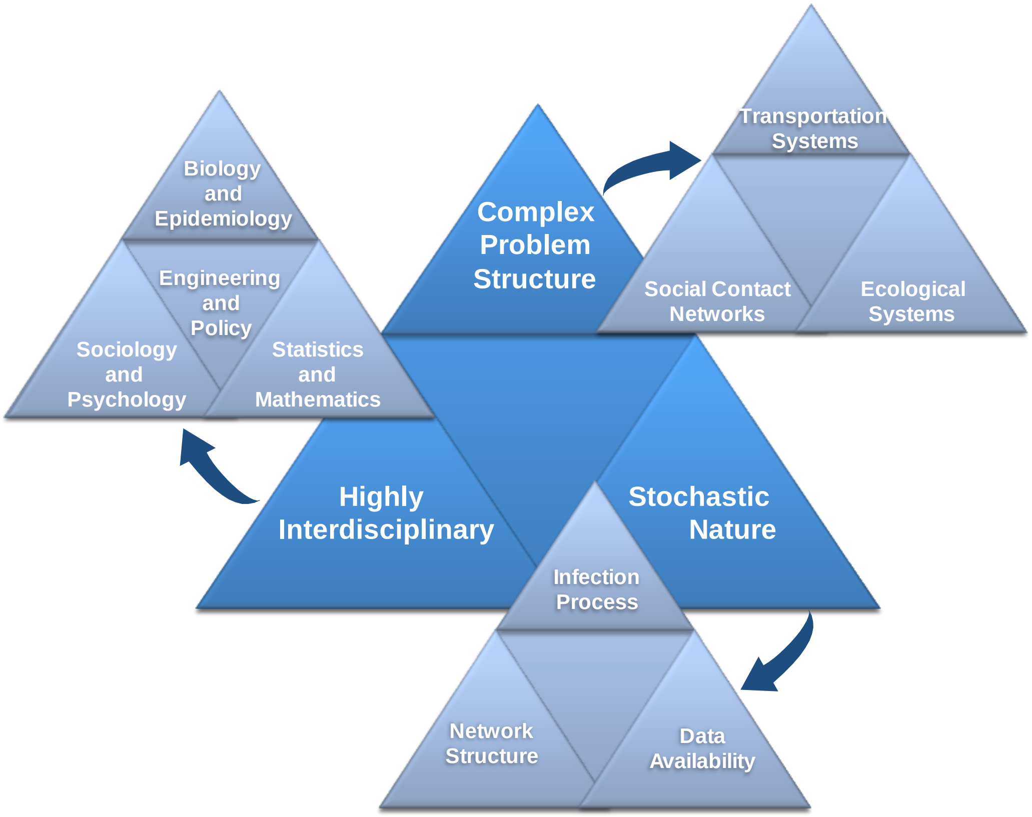


FIGURE 1-1: Research Challenges

The most common tool for modeling outbreak patterns is network theory. Network models can be used to model real world phenomenon across a broad spectrum of disciplines. Modeling the dispersal of infectious agents within and between regions lends itself to network analysis due to the natural structure of the environment on which it spreads. Human-to-human contacts have been extensively researched and defined under the title of social networks, while the field of transportation has existed as a network-based science for nearly a century. Methodology used in network based modeling and optimization problems often extends beyond the specific application in question, and can provide insight for new applications, such as those proposed in this research. Various related network-based problems are reviewed.

1.2 Related Applications

Beyond the realm of epidemiology, there are copious problems which involve the dispersal of "packets" throughout a network, where a "packet" is intended to represent a discrete item (germ, virus (computer or biological), unit of power (or lack thereof), chemical (biological warfare), currency, etc.). Examples of such problems are briefly defined below. Although each of these problems is interesting and deserving in its own right, most will not be expanded upon beyond this section.

1.2.1 Social Networks

Social networks have been used to model the spread of various psychological behaviors such as hysteria, depression, happiness and alcoholism (Fainzang, 1996; Benedict, 2007). Social network properties have also been exploited in the spread of computer viruses and mobile phone viruses, which can be modeled using internet or telecommunication networks. Newman (2002) explored the spread of computer viruses via a directed social network to define links over which the virus spreads. Wang (2009) modeled the mobility of mobile phone users in order to study the fundamental spreading patterns that characterize a mobile virus outbreak. Additionally, social network models have been applied in the study of obesity (Christakis, 2007).

1.2.2 Aquatic Networks

Another highly prevalent family of infectious diseases are waterborne, caused by pathogenic microorganisms which are directly transmitted when contaminated fresh water is consumed, and can be in the form of protozoa, viruses, or bacteria. There are countless waterborne diseases currently affecting the people around the globe (currently there is an outbreak of Cholera in Haiti). These types of diseases (most commonly diarrheal diseases) account for an estimated 4.1% of the total DALY global burden of disease, and are responsible for the deaths of 1.8 million people every year. It was estimated that 88% of that burden is attributable to unsafe water supply, sanitation and hygiene, and is mostly concentrated in children in developing countries (WHO). As these diseases rely on a contaminated water source, humans can become infected by consuming contaminated fresh water sources in nature or ingestion of a product in which contaminated water was used during production. A network can be used to model both these cases; representing either an aquatic network of streams, rivers, etc., or the logistics based distribution patterns of the contaminated product. However, the problem of modeling waterborne disease outbreaks is further complicated because the extent of the outbreak is dependent on the interaction between the network responsible for the contamination source (aquatic or logistics) and the human behavior network. This theme of co-dependent networks is also revealed in the spread of tick-borne diseases (i.e. Lyme), which is inherently a function of animal (i.e. deer) migratory patterns, as ticks require animal (sometimes human) hosts to survive and relocate, geographic spatial networks, and the corresponding ecological conditions.

In a related problem, the logistical distribution network can also be exploited to infer the most likely source by backtracking infection reports in a reverse engineering methodology. Such methodology might be desired to track food-borne outbreaks (i.e. salmonella) back to their source; or in the case of biological warfare in which dangerous toxins are released at one or more sources into regional infrastructure systems.

1.2.3 Power Networks

A problem which has received extensive research efforts, and shares many behavioral characteristics with disease dispersal is cascading failures in power networks. Cascading failures in power networks result when a disruption occurs in the network; a localized node/edge failure triggers the failure of successive parts of the system based on the dynamical redistribution of the flow on the network. This occurs when the load from the initial failure is absorbed by neighboring nodes, which are then pushed beyond their capacity so they become overloaded and fail, thereby further shifting their load onto other elements. There are various potential forms of disruption, which can be categorized as either random failures or targeted failures. In a random failure, randomly selected nodes/edges are removed from the network. For example broken branches falling on power lines during storms would constitute a random failure. In a targeted failure strategically chosen nodes/edges are removed from the network, such as the maximum capacity links, or the nodes which carry maximum loads. In a worse-case scenario this cascading effect can propagate through an entire system, resulting in total failure. The cascade stops when nodes are no longer being loaded beyond their capacity. This type of behavior is commonly seen in high voltage systems, where a single point of failure at a fully loaded or slightly overloaded system results in a sudden spike across all nodes of the system. This surge current can induce the already overloaded nodes into failure, setting off more overloads and thereby taking down the entire system in a very short time. Under certain conditions a large power grid can collapse after the failure of a single transformer. This contagion of overload follows paths of physical connections in the power grid; analogous to the contact requirement in disease spreading propagation. In addition, the network structure of power grids shares the scale free network property with social networks and many transportation networks; therefore both are susceptible to an exponential rate of failure.

This is exactly what happened in August 10, 1996 when a 1,300-megawatt electrical line in southern Oregon sagged in the summer heat, initiating a chain reaction that cut power to more than 4 million people in 11 Western States. This was also the likely behavior in August 14, 2003 when an initial disturbance in Ohio triggered the largest blackout in the US's history in which millions of people remained without electricity for as long as 15 hours (Crutitti, 2004). This type of cascading behavior parallels disease outbreaks, in which a critical initial case resulted in a devastating aftermath (before adequate intervention strategies could be implemented).

Targeted attack of telecommunications systems is also an integral issue due to the real-time communication dependency of so many major infrastructure systems in our modern world; as there is an inherent dependency between civil networks such as power and water distribution grids, gas transmission lines, transportation systems, and emergency response buildings, among others. For example the co-dependencies between a water, gas and power network are the following: water is required for production, cooling, and emissions reduction in both the power and gas networks; power is required for pump stations, lift stations, and control systems in the water network, and for compressors, storage, and control systems in the gas network; and gas is required for generators in the power network and heat in the water network. Similar interdependencies exist between various network systems which play a role in disease dispersal.

Cascading failures are also prevalent in economic systems (Sachs, 2008), again confounded by the interdependencies between many players in the financial market. Additional, but unrelated analysis that has explored interdependent networks in reference to the overlap of working memory and spatial attention networks (Agnati, 2007); as well as the overlap of crime and terrorist networks which assumes terrorists are also involved in local crime networks (Atkinson, 2009).

1.2.4 Transportation Networks

Similar methodology to that proposed in this research has potential applications beyond infectious disease spreading as well, such as information spreading in a social network. A further removed application that shares similar problem dynamics with contact disease transmission in a population is vehicle to vehicle wireless communication in a transportation network. One benefit of vehicle-to-vehicle communication is the ability to share real time traffic information directly between vehicles (passing vehicles can transmit travel times to one another gathered from previously traveled links). This information can then be used to update link travel times, and re-evaluate route costs. Given a set of vehicle trajectories and wireless communication properties (i.e. vehicle proximity requirements for information transmission), information sharing can be modeled similarly to disease transmission; aiding in the assessment and benefit of vehicle-to-vehicle wireless communication capabilities.

1.3 Research Goals and Objectives

The focus of this dissertation is predicting the spread of disease across network systems. Both contact-based and vector-borne diseases are explored. Contact-based diseases spread through direct human contact (i.e. influenza. sexually transmitted diseases), while vector-borne diseases spread to humans through the bite of an infected vector (dengue, malaria, yellow fever). The additional role of the vector alters the inherent nature of the infection process, thus requiring separate modeling tools to depict the process. Both types of infection processes are modeled, within the context of regional (social and spatial) and inter-regional (transportation) networks.

Mathematical models are proposed which use transportation systems (via transportation infrastructure networks), human activity patterns (via social networks), and ecological conditions (via spatial networks) to predict the local and regional dispersal of infectious diseases. Additionally, the prediction tools developed in this research use information (e.g. infection data) to infer specific outbreak scenarios; in contrast to the majority of epidemiological modeling tools which derive expected properties of an outbreak. The methodologies developed have the potential to be expanded in many directions for the problems proposed, as well as to other applications.

1.4 Dissertation Outline

The remainder of this dissertation is broken into five chapters. Chapter 2 includes an extensive literature review covering network structures (as they will be implemented in this work), and the leading epidemiological network models at both the microscopic (regional) and macroscopic (inter-regional) levels. Brief introductions for each of the proposed research problems are also included. Chapters 3-5, address each of the proposed problems in detail, including a problem description, solution methodology, sample application, numerical results, conclusions and future research direction. The proposed problems are introduced by chapter in order of problem scope.

Chapter 3 introduces a problem which uses available infection data to identify the path of infection (for a specific disease scenario) through a social contact network, by invoking the use of network based optimization algorithms. This problem can be thought of at the microscopic level because individuals are explicitly accounted for. Chapter 3 introduces a method for evaluating a social network which has been exposed to infection; improve prediction capabilities on future potential epidemic outbreak patterns, and aid in evaluation of potential intervention strategies, without running computationally intensive simulations. It also compliments a novel and related problem in phylodynamics, which uses genetic sequencing data to reconstruct the most likely infection spreading path.

Chapter 4 introduces a new application of the methodology introduced in chapter 3. The objective is to identify links in a transportation network responsible for spreading infection into new regions (spanning from a single source), using regional level infection data (e.g. when the disease arrived at a new location) and air traffic patterns. Like social networks, the inherent structure of a transportation system makes it an obvious candidate for network modeling tools. The new network structure is defined by nodes which represent regions (cites, states, countries), and links representing travel routes. This is a macroscopic application of the problem introduced in chapter 3, because individuals are no longer explicitly accounted for, though they are implicitly included in the model in travel volume. Additionally, because the actual contact level network is not included, this regional level model could be applied to a variety of infectious diseases, with the constraint that they originate at a single initial source.

Chapter 5 redirects the focus from contact-based diseases to vector-borne diseases, specifically exploring the role of international air travel in the spread of Dengue. Vector-borne diseases are transmitted to humans through the bite of an infected vector (i.e. mosquito), including dengue and malaria. Billions of people around the globe are exposed to these diseases annually, with millions of suspected infections. However, the role of the vector in the infection process inherently alters the spreading process (compared to human contact diseases), which must be addressed in prediction models. The objective of the problem is to quantify the relative risk posed by various international air travel routes for importing dengue infected passengers into susceptible regions. Various extensions of this model are proposed, including one for a multi-modal transportation network system.

Chapter 6 concludes this dissertation with an overview of the work presented, noting associated contributions and critiques of the research. Future directions for this research are discussed, including an example of interdependent network analysis.

CHAPTER 2: LITERATURE REVIEW

The highly interdisciplinary nature of this work demands an understanding of the state of the art modeling tools across various topics; including transportation engineering systems, ecological systems, infectious biological agents, human mobility and behavioral properties, mathematics, and statistics. A complete literature in just one of these fields is a daunting task on its own, and a comprehensive review is infeasible. Therefore the literature introduced in this chapter aims to provide the necessary background for the proposed research, and highlight the pieces of literature directly applicable to the problems at hand. The chapter begins with an overview of network structure and disease spreading behavior, introduces the state of the art models for predicting disease dispersal within networks of various size and structure, and presents various models for integrating network systems. Throughout this chapter the three main research problems addressed in this dissertation will be briefly introduced.

2.1 COMPLEX NETWORKS

Networks are commonly used to represent real world problems for mathematical modeling applications, and as previously mentioned, are well suited for disease spreading applications. In its most general form a network is composed of nodes and links (connecting these nodes). The links can be directed or undirected. The interpretation of the nodes and links depends on the specific application in question. The majority of networks used to model disease spreading fall under the umbrella of complex systems, characterized by diverse behaviors that emerge as a result of non-linear spatiotemporal interactions among a large number of components. [There are many additional types of network structures available to modelers; however the focus of this literature review is disease modeling on networks, not network theory; due to their applicability to the problem at hand only complex network structures will be reviewed in this section.]

For example a complex network can depict a transportation system representing air travel, where the airports correspond to nodes and the flights correspond to directed links (representing travel between airport pairs). If desired each link can be assigned a weight representing characteristics of the respective route such as travel volume, distance, etc. Transportation networks naturally lend themselves to network theory because system components clearly form a tangible network; which can be realistically modeled providing the necessary data (e.g. set of airports and travel routes and volumes).

Social contact patterns also form a complex network. Social networks are used to represent the interaction between individuals. In social networks nodes correspond to individuals, and contacts between the individuals define the links (which can be directed or undirected). [In disease network modeling the links are usually representative of infection spreading contacts and are therefore directed.] In contrast to transportation networks, social networks are difficult to accurately define because human interaction is complex and stochastic, and the required information is often unavailable. Even the most advanced activity based modeling tools make assumptions on human behavior. However it is still necessary to be able to model a realistic depiction of social interaction in a population to accurately predict disease spreading behavior. This is made possible due to the significant research that has focused on deriving the properties of social networks (e.g. size, degree distribution, etc.). These properties are often used to create a social network structure for modeling purposes. Similar methods are often used to model network systems other than social networks, when information on the network structure is unavailable.

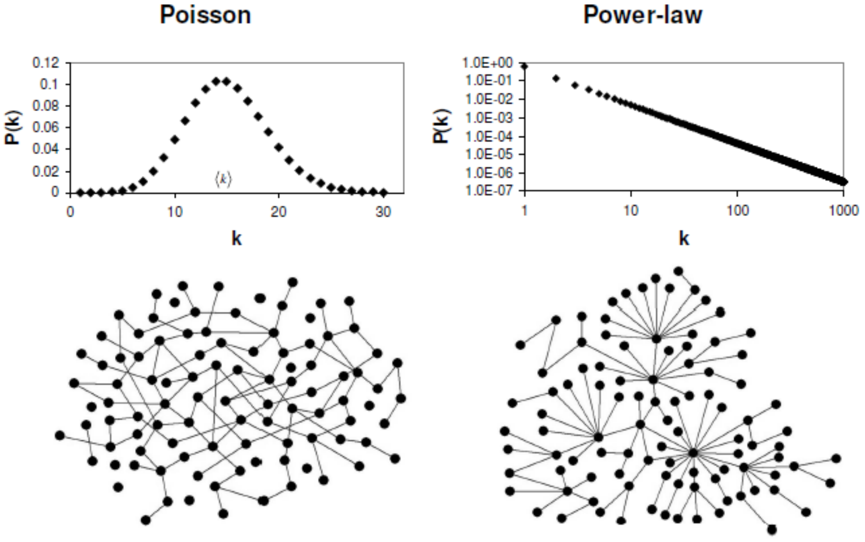


FIGURE 2-1: Network Structures and Distributions

Complex systems are scale invariant; where the structure of the system is similar regardless of the scale (example: fractals), demonstrate infinite susceptibility/response; a small change in the system conditions or parameters may lead to a huge change in the global behavior (examples: power grid, sand pile), and illustrate self-organization and emergence; in which a system can evolve itself into a particular structure based on interactions between the constituents, without any external influence, and a completely new property arises from the interactions between the different constituents of the system (example: ant colony).

There are two main categories of complex network structures that apply to disease spreading models because they are representative of the environments on which diseases disperse: 1) those with a peaked degree distribution, and 2) those with a power law degree distribution. An example of each of these networks is provided in Figure 2-1 (Thadakamalla, 2007). Network structures with a peaked degree distribution include random networks, such as that introduced by Erdős–Rényi (1959) which sets an edge between each pair of nodes with equal probability, independently of the other edges. In these networks most nodes have a degree near the average node degree, <*k*>. Networks with homogenous degree distributions are resilient to targeted attack because the probability of any node having an extremely high degree is low; therefore the probability of highly disrupting the network by removing the highest degree nodes is also low. Homogenous networks however react similarly to random failures (as they do to targeted attacks), because the degree of a randomly removed node is also likely to be close to the average node degree for the network. A random network (created with edge independence and equivalent edge existence probabilities) is inappropriate for modeling most real life phenomenon, including social networks, and additionally not likely to exhibit scale free properties, known to exist in many real world networks.

Small World Networks

Some of the more mathematical based modeling approaches for disease spreading are founded on the largely intuitive idea of early interruption of critical social contacts (World Health Organization, 2003), which is successful under the assumption that social contact networks can be represented as small world networks. [This same strategy would be effective for disease intervention in alternative networks with similar structural properties (i.e., hub and spoke air traffic).] The small world network idea arose is the 1960's when Stanley Milgram performed one of the first (very simple) quantitative studies of the structure of social networks. He concluded rather cavalierly that the average number of acquaintances separating any two people on the planet is six. This has since been labeled "six degrees of separation" (Guare 1990). Although Milgram's experiment was a bit unconventional, the result that two randomly chosen human beings can be connected by a short chain of intermediate acquaintances has been verified, and is known as the small world effect. This is a crucial result for human communication networks, specifically concerning disease spreading, which occurs via human-to-human contact.

The small world effect can be explained using a random graph. Given a population of size N, and assuming each person has on average z acquaintances (z is known as the coordination number), there are $Nz/2$ connections between people in the entire population. A random graph is constructed by taking N nodes, and adding $Nz/2$ edges between randomly chosen pairs of nodes to represent these connections. This represents a social network clearly displaying the small world effect. If a person A on the graph has on average z neighbors, then person A has on average z^2 second neighbors, and z^3 third neighbors, etc. The average number of degrees of separation, D, needed to reach all individuals on a network can be found by setting $z^D = N$. $D=logN/logz$, and this logarithmic increase in the number of degrees of separation with respect to the network size is typical of the small world effect. One other important property of small world networks develops from the idea that people's social circles tend to overlap, and therefore person A does not actually have z^2 second neighbors, since it is likely that many of those friends of friends are likely also friends of person A. This property is called clustering of networks. In a random graph the effect of clustering does not appear, and the probability that two acquaintances of person A know each other is the same as the probability that any two random people in the graph know each other. For clustering to be incorporated into the network, a clustering coefficient *C* is defined as the average fraction of pairs of neighbors of a node which are also neighbors of each other. This follows the idea that if two people know the same person, it is more likely they will know each other. In a graph where everyone is connected, $C=1$. In a random graph $C=z/N$. Reality lies somewhere in between.

In order to model realistic social networks both clustering effects and small world properties must be included. Watts and Strogatz (1998) developed a simple generation model that produces random graphs with small-world properties, including short average path lengths and high clustering. However this model still produces graphs that are homogeneous in degree, and does not account for hubs and the scale free properties in realistic networks. The Watts-Strogatz model seen in Figure 2-2 is created by forming a d-dimensional lattice with each site connected to its z nearest neighbors and then rewiring a small fraction of the links to new sites chosen at random, with probability p. These longer connections represent more distant acquaintances, while the z short connections represent daily acquaintances such as family or coworkers.

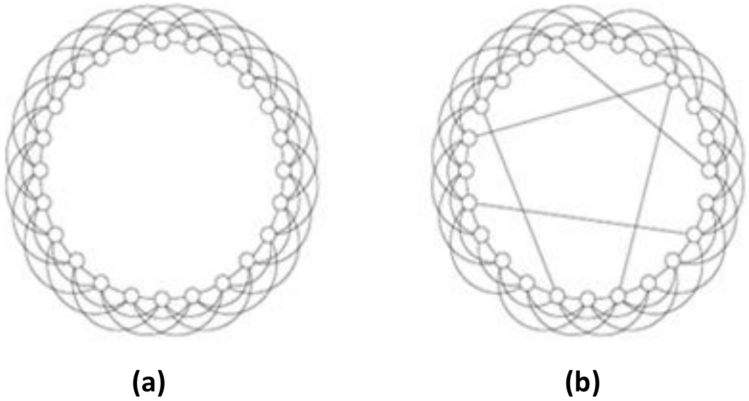


FIGURE 2-2: (a) SWN with each site connected to its z nearest neighbors and (b) an example of the Watts-Strogatz model with random rewiring between sites

Scale-Free Networks

The second category of network structures includes scale-free networks, identified by a power law degree distribution. Many real world networks can be represented as scale-free, including social networks, air travel, road maps, food chains, electric power grids, metabolite processing networks, neural networks, voter networks, telephone call graphs, WWW, movie actors, and military support logistics networks. These network structures are characterized by their heavier tailed degree distribution, where the majority of nodes have a very low degree and a select few nodes are highly inter-connected (example: hub and spoke network structure). This network structure results in an inherent robustness to random node removal, due to a low probability of randomly selecting one of the few highly connected nodes. However, scale-free networks are vulnerable to targeted attack because a few highly connected nodes are responsible for connecting a majority of the network. A model for generating random scale free networks was addressed by the Barabási–Albert model (1999) by incorporating growth and preferential attachment (however the clustering coefficient in this network structure is lower than that in the Watts and Strogatz model). Preferential attachment is the concept that a more connected node in the network is more likely to attract a newly incoming link, compared to a less connected node. This is related to the idea of social links in human contact networks. The Watts and Strogatz and Barabási–Albert models both contribute to the development of network representations of social contacts, (which display high clustering properties, and low average path lengths, and power law degree distributions), though neither is fully realistic.

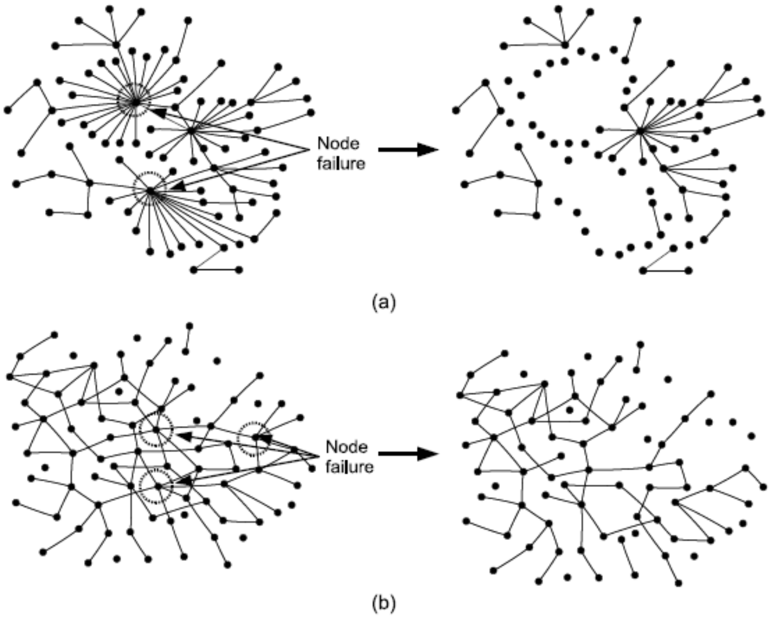


FIGURE 2-3: Example of Network Sensitivity

These homogenous and scale free network structures display significant variation in their resiliency to structural disruptions. An illustration of relationship between network structure and vulnerability is provided in Figure 2-3 (Thadakamalla, 2007), where the scale free network structure (a) is shown to be much more vulnerable to targeted attack, (i.e. removal of the two highest degree nodes), resulting in a highly disconnected network. This provides a very simplified representation of the effect of quarantining “highly connected” infected individuals in a social network or restricting traffic out of a transportation hub in an air travel network. The network structure (b) has the same set of nodes and same number of links as (a), however the alternative arrangement of the links results in a much less devastating result when the three highest degree nodes are removed. This more homogenous node degree distribution (b) mitigates the effect of targeted node removal. Figure 2-3 exposes the underlying relationship between network structure and robustness to (targeted) failure. This relationship between network structure and robustness to failure is used as motivation for many outbreak interdiction strategies. However, before interdiction strategies can be specified, models for predicting behavior of the infection process must be developed.

2.2 REGIONAL-LEVEL DISEASE PREDICTION MODELS

The introduction of mathematical models into the study epidemiology dates back to the early 20th century. In 1902, Ronald Ross was awarded the Nobel Prize in Medicine for his remarkable work on malaria, but his greatest achievement was likely the development of mathematical models for the study of its epidemiology (Ross, 1916). Over the years this basic model has been expanded upon in various directions.

The prediction models introduced in this dissertation can be applied to either I) contact-based diseases or II) vector-borne diseases. Vector-borne diseases (such as dengue and malaria) are transmitted to humans through the bite of an infected mosquito;

as such the traditional models for contact-based diseases (at the regional level) do not adhere. In addition, both models will be explored on a i) microscopic (regional) scale (modeling individuals explicitly) and ii) macroscopic (inter-regional) scale (modeling regions explicitly). Social networks are the main network structure used in microscopic models (for contact-based diseases); in macroscopic models the network structure is defined by the inter-regional transportation systems. Microscopic level models for contact-based diseases are introduced first.

2.2.1 Contact-Based Diseases

The spread of contact-based infectious diseases is an inherently stochastic process. Because of the exponential nature of disease spread, real time control and prediction methods present a huge challenge, due to the dynamic and stochastic nature of the problem combined with imperfect information. A wide number of diseases are primarily spread through human interaction. We usually think of diseases as being spread through human populations between infective (those carrying the disease) and susceptible individuals (and those who do not yet have the disease but can catch it). For those diseases that spread through human to human contact, the pattern of disease spread can be modeled as a social network, where individuals are represented as nodes, and contact between individuals are represented as edges. The rate and pattern of the disease spreading process through a network is dependent on both the parameters of the disease (infectious period, level of contagiousness, etc.) and the fundamental structure of the network. The majority of research in the field of epidemiology focuses on development and implementation of agent based simulations to predict average disease spreading behavior and characteristics. At this point there is a lack of research focusing on disaggregate, real-time predictive and preventive measures for specific spreading scenarios.

Disease Spreading Properties

The reproductive ratio is a variable commonly associated with disease spreading. The reproductive ratio (R_0) can be thought of as the total number of new cases resulting from a single infected individual. It is defined as: $R_0 = \frac{\beta * S}{\upsilon + \delta}$; where β is the average rate at which infected individuals have contact with randomly chosen individuals of all states (infected, susceptible or immune), S is the size of the susceptible pool in the population, and ($\upsilon + \delta$) is the average rate at which infected individuals recover and acquire immunity, δ, (or die, υ). If $R_0 \geq 1$, then the rate at which individuals are becoming infected is higher than the rate that individuals are recovering, and therefore in theory a small outbreak could expand and become a large scale epidemic (though this is not guaranteed); and if $R_0 \leq 1$, then individuals are recovering faster than getting infected so the disease would die out on its own. [Note: An outbreak differs from

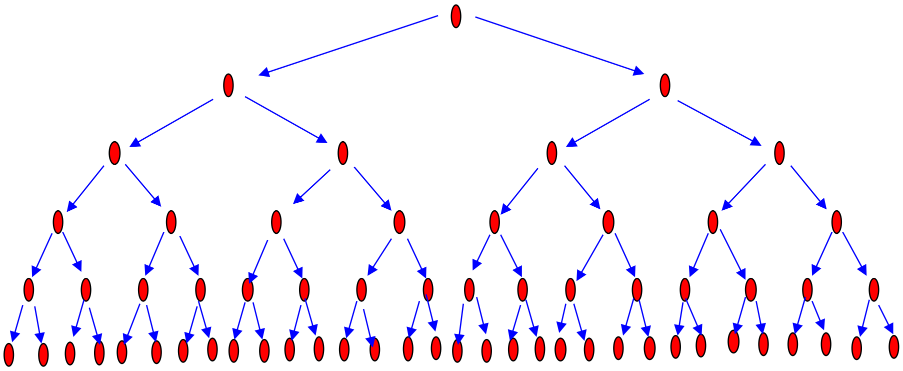


FIGURE 2-4: Example $R_0 = 2$

an epidemic because it is defined as casually connected clusters of cases that die out before spreading to the population at large, where as an epidemic results in population wide incidence of the disease. An outbreak is therefore determined by the spontaneous dying out of the infection, where as an epidemic is limited only by the size of the population in which it spreads.] By definition, the total number of expected cases of a disease should increase by R_0 for every generation of the infection. The exponential nature of disease spread is clear from the simplified example in Figure 2-4, where $R_0=2$ new cases generated per existing case. While R_0 may serve as a good intuitive explanation for whether a disease will spark a full scale epidemic, it has been shown that R_0 estimates have in the past not accurately represented disease spreading when extrapolated to a population at large. This may be due to two reasons:

i. R_0 values are generated based on the premise of fully mixed epidemiological models - the assumption is that all individuals in a group are equally likely to become infected. This is often not the case in reality.

ii. Estimates may be based on transmission data in closed settings such as hospitals or crowded apartment buildings with usually high rates of contact. When an estimated R_0 value is extrapolated to the broader community it may result in spurious estimates

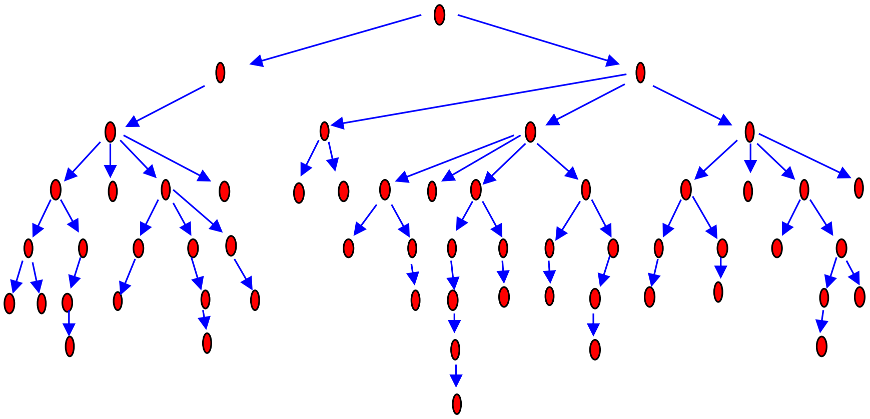


FIGURE 2-5: Example of R_0 Varying over time

For these reasons it is important to consider R_0 as a distribution of possible values, depending on the setting in which the disease is spreading. The effect of varying R_0 values can be seen in Figure 2-5. Another problem with R_0 is that a single R_0 value can result in vastly different epidemiological outcomes, depending on the contact patterns in a community.

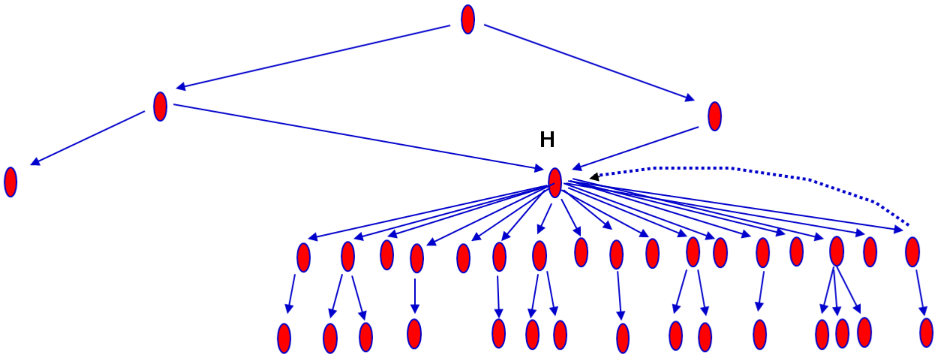


FIGURE 2-6: Example of a Super-spreader

Many infectious diseases vary from the standard fully mixed models, and often exhibit heterogeneity in transmission efficiency, where certain individuals are responsible for a large proportion of the transmission events. These individuals may be referred to as "super-spreaders". There is a large difference between the situation where all individuals share typical contact patters, and the one in which most individuals pass the disease on to zero or one other individual, but a few individuals pass it along to dozens. However, in both these cases the R_0 value could be the same. An example of a super-spreader can be seen in Figure 2-6.

Recent advances in disease modeling have addressed these issues. Haydon *et al.* (2003) develop a novel parameter–free method that permits direct estimation of the history of transmission events recoverable from detailed observation of a particular epidemic. From these reconstructed 'epidemic trees', dynamic case–reproduction ratios can be estimated directly, $R_0(t)$. To construct the epidemic tree they developed an algorithm that generated a putative source of infection (referred to as a 'parent') for each Infected Property (IP) in the following way: when the parent was known from contact-tracing, it was always the assumed infection source. When there was no contact-tracing information available we assumed that the parent itself must have been infected at least T days prior to the infection date of the daughter and extant (not culled) on or before the day of the daughter's infection. Subject to these conditions, the adopted parent was a selected IP from a 'candidate' list, located within a certain distance of the daughter (50 km was chosen as a compromise between an exhaustive candidate list and computational expediency). They adopted three rules for selection of parents from the list of possible candidates: (i) selected parents were simply closest to the daughter (i.e. a single tree was constructed deterministically), (ii) they were selected from the candidate list with equal probability; or (iii) parents were selected from the candidate list with probability inversely proportional to the distance from the daughter IP. For (ii) and (iii) 500 trees were created and average properties were analyzed. Numbers of daughters rising from each parent can be directly counted over the epidemic tree, and $R_0(t)$ estimated by averaging these values within a tree over different time intervals and geographical regions. They apply this method to data from the 2001 foot–and–mouth disease outbreak in the UK (Haydon, 2003). In addition to introducing a novel approach for deriving the reproductive ratio of a disease as a dynamic variable, this work incorporates a methodology to predict the spatial path of infection. While a simple epidemic tree construction algorithm is used here, this idea of using infection data to construct the most likely path of transmission is a highlighted topic of this research.

A significant advancement in disease modeling replaces the “fully mixed” model where the susceptible individuals with whom an infected person comes in contact with are chosen at random and with equal probability from the entire population, disease propagation can instead be modeled on a “contact network”, which only allows infection to occur between an infected and susceptible individual who are connected by an edge in the network. Second, the number of contacts (edges) each person (node) has may vary by applying different degree distributions to the network. Lastly, the probability of disease-causing contact between pairs of individuals may vary. These applications can capture the interactions that underlie the spread of diseases.

Transmissibility

As mentioned, one major assumption in most disease spreading models is that the probability of infection between any two individuals is the same for all pairs of individuals. Additional work by Newman (2002) has further extended such models by allowing the probability of disease causing contacts to vary between connected pairs of individuals, so some have higher probability of disease transmission than others. This is accomplished by introducing the "transmissibility", *T*, of the disease. Using applications from percolation theory in physics Newman derive analytical expressions to characterize disease spreading in networks as a function of transmissibility, *T*. *T* can be used as an alternative to the R_0 introduced previously.

Transmissibility is the average probability that an infected individual will transmit the disease to a susceptible individual with whom they have contact. *T* is a function of r_{ij}, the average rate of disease causing contacts between an infective individual (*i*) and susceptible individual (*j*), as well as τ_{i}, the amount of time the infective individual remains infective. *T* therefore encapsulates various attributes of the disease including the rate at which contacts take place, the likelihood that a contact will lead to transmission, the duration of the infectious period, and the susceptibility of individuals to the disease. If both r_{ij} and τ_{i} are iid random variables, then *T* is just the average of T_{ij} over the distributions *P(r)* and *P(τ)*, and $T_{ij}=1-(1-r_{ij})^{\tau\mathrm{i}}$. For any outbreak of the disease beginning with a single infected individual, and spreading across the network, we "mark" each edge in the graph on which the disease is transmitted, which happens with probability *T*. The ultimate size of the outbreak would be the size of the cluster of vertices that can be reached from the initial vertex by traversing only occupied edges. Therefore the model is equivalent to a bond percolation model with bond occupation probability *T* on the graph representing the community, and unlike R_o; *T* can be extrapolated from one location to another even if the contact patterns are quite different. The percolation model was first formulated in this manner by Grassberger (1983) for the case of uniform *r* and *τ*, and by

Warren *et. al.* for the non-uniform case (2002). Newman uses similar applications of the percolation problem on networks with arbitrary degree distributions to derive analytical expressions for the size of an outbreak, presence of an epidemic, and size of the epidemic, all as a function of the transmissibility. Probability generating functions are defined for the degree distribution, and used in solving for the average behavior on random graphs.

2.2.1.1 Probabilistic Models

In efforts to capture the behaviors described above, the current disease models span from extremely generalized and simplified mathematical functions, to increasingly in-depth stochastic agent based simulation tools. The simplified models may not be able to capture certain behavioral aspects of the dynamics of disease spreading because they lack certain details about the network structure and disease characteristics, while the more recent, and most comprehensive models may be the most realistic, they are also often too complicated to reproduce when an urgent situation arises. An ideal model would be the most simplified version that can be practically implemented in a timely manner, but is still complex enough to capture the realistic dynamic behavior of the disease. All of these models however fall under the category of probabilistic models, which predict expected outbreak behavior. These models do not account for real-time information, which would be a desirable asset during the onset of a potential epidemic.

2.2.1.1.1 Static Case: Mathematical Models for Disease Spreading

For modeling disease propagation through a contact network, the Susceptible-Infected-Removed (SIR) model is often implemented. The original and simplest SIR model was proposed by Lowell Reed and Wade Hampton Frost in the 1920's (though never published) and is as follows: A population of N individuals is divided into three states: susceptible (S), infective (I) and removed (R). Infective individuals come in contact with other individuals at a rate β, and recover and acquire immunity (or die) at a removal rate α. For a large population N, the proportion of the population in each of the three states at a given time t can be represented by a set of differential equations.

$$\frac{d}{dt}S(t) = S(t)\big(1 - \beta * I(t)\big) \tag{2-1}$$

$$\frac{d}{dt}I(t) = (\beta * S(t) + \big(1 - \alpha)I(t)\big) \tag{2-2}$$

$$\frac{d}{dt}R(t) = R(t) + \alpha * I(t) \tag{2-3}$$

$$S(t) + I(t) + R(t) = 1 \tag{2-4}$$

(The SIR model has an additional variation, the SEIR model, where the "E" represents an infectious period where an individual is infected but not yet symptomatic. The sequential nature of the states is represented in Figure 2-7.)

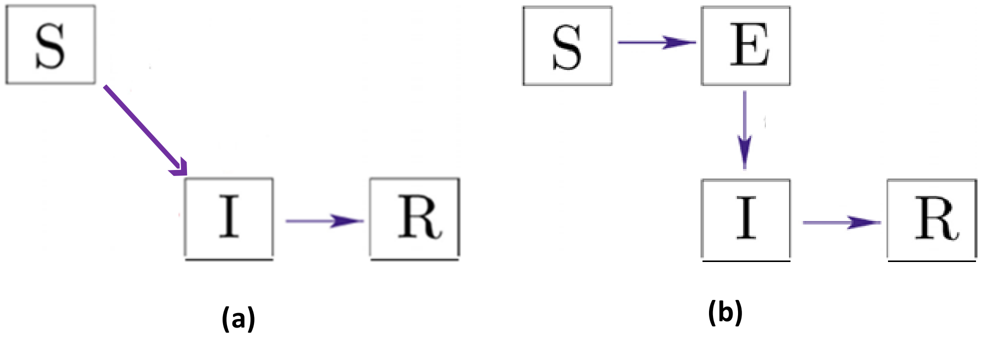


FIGURE 2-7: (a) Three stage S-I-R Model and (b) Four stage S-E-I-R model

The SIR model is fairly straight forward; however there are some faults with this model:

i. The biggest issue with the traditional SIR model is that the network used is traditionally a “fully mixed” model where contact between an infectious and susceptible person is random, with the same probability of contact for any two individuals in the network. Therefore disease does not propagate through the network like it would in a traditional social setting.

ii. Another issue arises because the parameters α and β are constants, but in reality they should vary throughout the time course of an epidemic. This is because as a disease arises, public health practices change and the rate of infection, β, should decrease over generations of the disease, while the removal rate, α, should increase because the amount of time to diagnose and treat the disease should reduce.

These issues are addressed in some of the more advances SIR models, and in alternative modeling tools such as agent based simulation.

2.2.1.1.2 Dynamic Case: Agent based Simulation models

Agent based simulation is another popular modeling tool in the epidemiological literature. Simulation is used to replicate possible spreading scenarios, predict average spreading behavior, and analyze various intervention strategies. Large scale simulation models are computationally taxing, and require known network structure and various parameter specifications including population characteristics and specific disease parameters.

By simulating disease spread on various small world networks Meyers *et. al*. (2005) found that the physical network structure, in addition to the disease parameters, plays a vital role in the propagation of diseases. Meyers found it is likely that different communities with similar contact patterns will have very diverse experiences with the disease, some resulting in only small outbreaks, while some resulting in a full scale epidemic. This implies that the standard SIR model (which assumes a Poisson distribution) cannot be generalized to arbitrary degree distributions, but that variation in transmission probability as a function of activity and realistic network connectivity are integral issues when modeling disease spreading. Small and Tse (2005) experimented with related disease parameters such as the probability of transmission (they considered two different values, one for long range connections and one for short range). They found that there is a breaking point at which the transmission probability between an infectious and susceptible person separated geographically is high enough to result in an uncontrolled outbreak. This type of simulation captures the overall number and distribution of the infected population, but not the temporal progression.

One of the more complex agent based simulations that exist is Epidemiological Simulation System (EpiSims). In 2004 Models of Infectious Disease Agent Study (MIDAS) was established by the National Institute for General Medical Sciences, part of the U.S. National Institutes of Health. MIDAS is a collaboration of research and informatics groups to develop computational models of the interactions between infectious agents and their hosts, disease spread, prediction systems, and response strategies. As part of this research effort EpiSims was developed, which is an in depth simulation of disease dynamics and includes a realistic network structure created through the use of population synthesis, activity assignment, location choice and travel time information (Eubank, 2005). EpiSims was used to evaluate various potentially feasible intervention strategies for influenza, such as quarantine, isolation, school closures, community social distancing, workplace social distancing, and also pharmaceutical intervention such as antiviral treatment. One simulation was conducted on a population similar in size to Chicago, and it was found that timely implementation of targeted antiviral treatment by household, along with social distancing could have a substantial effect on the illness attack rate. However, due to a lack of data, additional research is recommended to learn more about the sources of transmission and effectiveness of social distancing measures **(**Halloran *et al*, 2008).

2.2.1.2 Case-Specific Prediction Models

The literature reviewed so far details probabilistic models used to predict the expected spreading behavior of an infectious disease among a group of connected individuals. There is an obvious gap in the literature for scenario specific disease prediction models. These models would provide many of the same benefits as the probabilistic models, such as aiding in strategic intervention planning and providing insight to potential future outbreak conditions; coupled with the added benefit of being tailored to specific case studies, providing a much more acute level of analysis, and perhaps requiring less computational effort than the large scale agent based simulations. This approach to epidemiology is most prevalent in the field of microbiology.

2.2.1.2.1 Inferring Disease Spreading Patterns: A take on Phylodynamics

In the field of biology, *phylogenetics* is the study of evolutionary relatedness among various groups of organisms (i.e. species, populations, viruses), discovered through molecular sequencing data. Viruses are ideal candidates for studying their spatial and temporal dynamics through the use of phylogenetics because their rate of mutation is fixed (due to their rapid rate of nucleotide substitution), and therefore the branching structure of virus phylogenies provides a unique insight into their evolutionary patterns. It is important to note there is inevitable stochasticity in these predictions. The identification of these evolutionary patterns is a difficult problem; and one of the most advanced models for doing so is BEAST: a software architecture for Bayesian analysis of molecular sequences related by an evolutionary tree. This software incorporates a large number of popular stochastic models of sequence evolution and tree-based models suitable for both within- and between-species sequence data are implemented. (Drummond, 2007)

A related science, *phylogeography*, is the study of tracing the historical process responsible for the current geographic distributions of individuals, explicitly focusing on biogeography, rather than just the population genetics alone. Specifically, gene genealogy is interpreted to infer historical migration patterns and population expansion. Intuitively, similar methodology can be applied to tracking the geographic distribution of viruses. Methodologies borrowed from both phylogenetics and phylogeography provide valuable tools for predicting the spatial and temporal infection spreading patterns of a given virus. Given the recent improvements and availability of genetic sequencing data, this is a commonly researched problem by microbiologists and epidemiologists alike.

The main focus of phylodynamics is on developing statistical models to enable the reconstruction of timed viral dispersal patterns. In the most advanced models, phylogenetic uncertainty is accommodated using Standard Markov model inference is extended with a stochastic search variable selection procedure that identifies the parsimonious descriptions of the diffusion process. Using such models, reconstruction of the H1N1 Virus dispersal is presented by Lemey *et. al* (2009). In the same paper they propose priors that can incorporate geographical sampling distributions or characterize alternative hypotheses about the spatial dynamics. (Lemey, 2009). A similar problem is presented by Wallace (2007), which uses phylogeography of H5N1 genetic sequences to analytically infer the geographic history of the H5N1 virus's migration.

Cottam *et. al* (2008) combines epidemiological data that relate to the timing of infection and infectiousness, with genetic data that show the genetic relatedness of pathogens isolated from infected individuals into a maximum-likelihood approach to infer probable transmission trees. This is accomplished by first enumerating all possible evolutionary trees, then assigning posterior probabilities based on specifics of the respective virus' mutation rates. Additionally, the infection trees only include locations where samples were available; and there is no proposed method for inferring information or scenarios accounting to locations without sample.

The previous methods use statistical properties derived from the virus mutation process to reproduce the most likely infection spreading scenarios, where the scenario probability is calculated *a posteriori*. A novel approach to reconstruct the spatiotemporal dynamics of outbreaks from sequence data was presented by Jombart (2009), tracing the path of disease by using genetic sequencing data, ancestries are inferred directly between strains of an outbreak using their genotype and collection date, rather than through the reconstruction of most recent common ancestors (MRCAs) as in phylogenetics. The fundamental innovation of this approach is to seek ancestors directly from the sampled strains. The authors apply this method to track the 2009 H1N1 pandemic. The "infectious" links are selected such that the number of mutations between nodes is minimized. The results are compatible with current epidemiological understanding of the 2009 H1N1 pandemic, while providing a much finer picture of the spatiotemporal dynamics. This application chosen was highly successful because A/H1N1 was the first human pathogen routinely genotyped from the beginning of its spread. While the results highlight how much additional epidemiological information can be gathered from genetic monitoring of a disease outbreak, this complete set of data is almost never available. Problem I proposed in this research differs from Jombart's model in that our objective is to find the maximum probability spanning tree within a social network where the edge weights are a function of temporal infection data and transmission probabilities rather than genetic sequencing data. In addition, an extension of the initial model is proposed to account for missing information, while Jombart's model seeks ancestors only accounting for sampled strains, thus likely over-looking many infected regions.

2.2.1.3 Problem I: Inferring infection spreading links in a social contact network

The tackled research problem of tracking viruses through space and time using genetic sequencing data in combination with infection reports (spatial and temporal) serves as the main motivation for this work. However, more often than not the required genetic data (and mutation based statistical properties) is unavailable. For the problems proposed in chapter 3 available infection reports are used to accomplish the same goal, inferring the spatiotemporal path of infection. This methodology has two analogous applications that will be explored:

i. *Tracking infection patterns through human social networks*

ii. *Tracking infection patterns through regional transportation networks*

The objective of Problem I is to predict the infection spreading pattern of a specific disease scenario through a social contact network. This research will serve as a way to evaluate a social network which has been exposed to infection. In this problem setting a disease is already present in the network; and real time information is provided for the infected nodes (i.e. the identity of the infected individual and when they were infected).

Problem I uses available infection data and network structure properties to infer the most likely path of infection. This varies from the reviewed research because it is utilizes optimization methods to reconstruct a tree using available infection data, rather than enumeration followed by *a posteriori* analysis. Additional problems beyond disease spreading can also be modeled with this methodology, such as information spreading in a social network.

2.2.2 Vector-borne Diseases

The bulk of epidemiological literature focuses on the family of diseases referred to as contact-based, which are transmitted between humans through direct contact. An additional family of diseases is arboviruses, which are transmitted from person to person through the bite of an infected mosquito, with humans serving as the main viral host (and reservoir). The geographic establishment of an arboviral disease, (i.e. dengue) is thought to be limited purely by the spread of its principal vector mosquito species (e.g. *Ae. aegypti* and *Ae. Albopictus*), therefore transmission reduction requires local control efforts to rid of mosquito populations. For many such diseases, the principle vector species have proven to be highly adaptable to human habitation. Population growth, urbanization, deforestation, poor housing, inadequate sewage and waste management systems, lack of reliable water systems, and increased movement of people, pathogens, and mosquitoes contribute to continued geographic spread, increased suitability for vector species establishment, and increased incidence of the disease (Gubler, 2001), and as a result, the global spread of the vectors can be difficult to contain (WHO, 2010).

As stated, the main strategy in the prevention and control of dengue within a region is "source reduction", or prevention of breeding places, specifically preventing the mosquito (*Aedes aegypti*) that transmits dengue from breeding inside and in the vicinity of homes. This can be accomplished by preventing existing water collections from becoming places for breeding of *A. aegypti* by draining out water from various containers, by regular changing of water and other items or, in the case of unused items, by discarding/destroying them. Since the mosquito does not travel far, "house cleaning" by all members of a community will ensure that no breeding places exist, preventing dengue from occurring. Additionally the spread of dengue from a patient to others must be limited by protecting the patient from contact with mosquitoes, which would bite the patient, thereby get infected and further spread it to others. This can be achieved by ensuring that the patient sleeps under a bed-net, using effective mosquito repellents are used where the patient is being provided care. This will prevent the mosquito from biting the patient and from getting infected and spreading it to others.

Clearly, modeling the spread of vector-borne diseases through a human population is not an analogous problem to modeling the spread of contact-based diseases, as the infection process is inherently dependent on the additional role of the vector. These spreading paths do not have a clearly defined structure, and therefore network-based mathematical modeling is not a technique currently employed by researchers working to predict how vector-borne diseases disperse within a region.

The only use of mathematical modeling (in the literature) to quantify the risk estimates for acquiring dengue (within a region) was proposed by Massad and Wilder-Smith (2009). Massads' model is intended to evaluate the risk of infection at a specific destination as a function of human population size, the number of infected mosquitoes, and estimated parameters for the mosquitoes biting rate and the probability that an infectious mosquito will infect a susceptible human. They first calculate the force of infection, defined as the per capita number of new cases per time; a function of the total human population, the number of infected mosquitoes, the mosquito biting rate and the probability that an infectious mosquito will bite a susceptible human. The probability of an individual acquiring dengue, was calculated based on the time of arrival, and duration of stay, where the numerator represents the total number of new infections occurring during the time of stay, and the denominator is the population size during that period. The results included probabilities of acquiring infection for different combinations of arrival periods and stay durations. The results found that arrival in high and low season results in drastically different infection probabilities. While this model is one of the first mathematical models aimed at calculating infection risk for travelers, the methodology varies from the proposed model in this work in many ways. This model does not account for travel patterns, or species distribution data in its prediction; and lacks quantitative validation from infection data. This is an interesting problem, but one outside the scope of this research. However, by aggregating regional infection data, the inter-regional dispersal of vector-borne diseases does lend itself to network analysis.

2.3 GLOBAL-LEVEL DISEASE PREDICTION MODELS

From the literature reviewed so far it is clear there is generous work on predicting the spreading behavior of contact-based diseases at a regional scale. The next logical step is to predict disease spreading behavior on an inter-regional scale. History has exemplified the significant role of modern transportation in furthering the spread of diseases across cities, states, countries and continents. Today infected humans have the potential to carry viruses into new geographical areas through air travel. Additionally, a substantial rise in international air traffic has increased the potential for virus dispersal into previously unoccupied regions. This burgeoning risk serves as the main motivation for this research.

Currently contact based diseases constitute a significant portion of the global-level disease modeling literature; however the models used to predict spreading behavior of contact-based diseases are not directly applicable to modeling diseases that spread through alternative infectious agents (i.e. vectors). This section will introduce the current models for evaluating the impact of air travel on the global spread of contact-based diseases, followed by the relevant modeling techniques for predicting the global dispersal of vector-borne diseases.

2.3.1 Contact-Based Diseases

The recent global outbreaks of SARS (2003), Avian Flu (2004, 2005, 2006, and 2007) and Swine flu (2009) among others have motivated methodological advancements for integrating inter-regional transportation patterns into previously regional-level modeling tools. The most common approach to modeling the global spread of human contact-based diseases is extending the (regional-level) mathematical and simulation based compartmental (SIR) model to incorporate regional inflows and outflows of individuals based on travel data. These models must often make various simplifying assumptions in order to model any large scale application.

The current probabilistic models which couple local infection patterns (using SIR) and air travel are based on the original model proposed by Rvachev and Longini (1985), with an extension focusing on prediction by (Longini, 1986). Hufnagel, *et. al* (2004) propose one of the latest advances derived from this type of probabilistic model for predicting global level epidemics. The authors implement a microscopic description of traveling individuals via a stochastic simulation, where the SEIR model to replicate local infection dynamics (using parameters specific to the 2002 SARS outbreak originating in China), is coupled with a stochastic dispersal of individuals between cities to account for the potential global spread of an epidemic. The inter-city dispersal is defined by a transition probability matrix, which is the probability an individual will travel between cities; and is a function of the travel volume (specifically the proportion of individuals on a specific route out of a given airport, compared with the total outgoing travel volume of that airport), and the typical amount of time an individual spends at a given city. This coupled model provides a tool for predicting global epidemic patterns by allowing the dispersal of infected individuals into previously uninfected regions. Simulations are run exploring various rates of transition rates (between cities), and various interdiction scenarios (e.g. individual travel restrictions, city-based travel restrictions, and route-based restrictions) for hypothetical outbreaks, and actual global travel patterns. The authors explore interdiction strategies including vaccination requirements (as the probability of having to vaccinate a certain percentage of the population to prevent an outbreak) and travel restrictions. Tor the interdiction scenarios explored the results suggest isolating cities (restricting travel out of the largest cities) and imposing individual travel restrictions is more effective than restricting travel on the highest traveled routes. This model is purely speculative, and does not account for any actual infection reports or known outbreak scenarios.

A related simulation based applications was presented by Grais (2004), though the concentration was at the country level. Coupling the standard deterministic compartmentalized SEIR model with U.S transportation data, Grais sought to identify whether U.S. air travel patterns lead to a better forecast of epidemics. This simulation-based model allows travel of susceptible and latent individuals to move between cities, but not infected ones. Various other simplifying assumptions are made. The model was run for two cases: 1) with air travel and 2) without air travel. The predicted forecasts for both cases were compared to various sources of influenza data (observed epidemic set, peak and end) to try and identify whether or not accounting for air travel improves the model's epidemic forecasts. The authors found air travel does appear to play a role in the spread of influenza, however model restrictions prevented a more complete and realistic analysis.

Around the same time Brownstein *et al* (2006) provided the first empirical evidence for the role of airline travel in long-range dissemination of influenza, by assessing the role of airline volume on the yearly inter-regional spread of influenza in the United States. Using weekly influenza and pneumonia mortality from the Centers for Disease Control and Prevention and air travel volumes, they measured the inter-regional spread and timing of influenza in the United States for nine seasons, from 1996 to 2005. The goal was to determine if the observed seasonal influenza peaks may have been influenced by air travel. They modeled the response of inter-regional spread of seasonal influenza to fluctuations in domestic air volume; and investigated the effect of international airline travel on the absolute timing of nationwide seasonal peaks. Regression models (using yearly travel fluctuations) fit to the seasonal infection data suggested an important influence of international air travel on the absolute timing of influenza introduction, as well as an influence of domestic air travel on the rate of inter-regional influenza spread within the US.

Cooper (2006) found somewhat contrasting results. In a paper analyzing the probabilistic effect of various interdiction strategies, they found on average, restrictions on air travel (e.g. travel to and from infected cities) are likely to be of surprisingly little value in delaying epidemics, unless almost all travel ceases very soon after epidemics are detected. Instead interventions to reduce local transmission of influenza (e.g. isolation, behavioral changes, antiviral use, etc.) are likely to be more effective at reducing the rate of global spread and less vulnerable to implementation delays than air travel restrictions. The model used was an extension of the coupled epidemic transmission model implemented by Hufnagel (2004) to simulate the international spread of avian influenza. The model parameters where decided by those that best fit the 1968/69 influenza pandemic including seasonal variability, R_0, transmissibility between tropical and temperate regions, the distribution of the infectious period, the initial proportion of susceptible individuals, first city infected, and the date of initial infection.

As was the case for regional-level disease prediction, both simulation and analytical models are implemented for the global prediction. Colizza, *et al.* (2006), developed a probabilistic mathematical model to explore the role of the stochastic nature of disease transmission, international travel flows, outbreak initial conditions and network structure on the statistical properties of global epidemic patterns. The authors use data from the worldwide airport network (WAN) and the International Air Transportation Association database (IATA) accounting for 99% of the worldwide air traffic, as well as census data for city populations, and explicitly calculate the disease evolution in all major urban areas connected by the global air travel network. The model developed is an integration of stochastic compartmental SIR models (modeling the infection dynamics for each city), coupled together by the use of a stochastic transport operator, to describe the movements of individuals between cities. The dynamics of the disease are calculated, and the spatiotemporal pattern is evaluated as a function of network structure. The authors conclude the heterogeneous air-transportation network properties play a significant role in the global dispersion of disease, and that large scale mathematical models can provide quantitative measurements on the predictability of epidemic patterns. A more detailed explanation of the model and further analysis of global outbreaks and the predictability of an epidemic to the structure of the transportation network are provided in Colizza, *et al.* (2006).

Additionally Balcan *et. al.* (2009) introduced a similar model to incorporate multiple scales of human mobility: i) small scale (intercity) communting flows estimated using a gravity model in conjunction and ii) long range traffic estimated using international air travel patterns, into a integrated worldwide structured metapopulation epidemic agent based simulation model. The model is intended to evaluate the role of multiscale human mobility in the infection spreading process. The authors found that short-range communting flows only result in small variations with respect to the base case which only considers airline traffic, even though they are estimated to be an order of magnitude larger than airline flows. The commuting flows do however synchronize the infection process among subpopulations within close proximity. This work is introduced for two reasons: i) it is an excellent example of a large scale epidemic prediction model built around a transportation-based system, and ii) it provides an example of a layered computational approach, where a unified multiscale network is used to represent interdependent transportation systems, a reoccurring theme of this dissertation.

2.3.1.1 Problem II: Inferring infection spreading links in a transportation network

The second problem proposed in this research is an extension of problem I, though applied at the inter-regional level. This objective of this macroscopic version of the problem is to identify the travel routes which are the most likely responsible for transporting infected individuals to previously unexposed regions. Input for this model includes transportation network properties (travel routes and volumes) and temporal regional infection data; and outputs a spanning tree representing the most likely inter-regional travel routes which corresponds to the outbreak data. The assumptions, link probabilities and constraints will vary from problem I, however a similar solution methodology is implemented. Due to the availability of data the application chosen for this research is the Swine Flu outbreak (2009) within the U.S. More detail on this model is presented in Chapter 4.

A potential extension of this problem is to infer the most likely *source* of infection, using information on current infected/contaminated sites. Example applications for this problem include back tracking food borne outbreaks (i.e. salmonella, *e coli*, etc.) to their respective source (a manufacturer, warehouse, distribution center, etc.), perhaps through a network structure representing a logistical distribution systems. This is a problem which will be further explored in future research.

2.3.2 Vector-borne Diseases

In addition to the known impact of travel on contact-based diseases, travel is suspected to be a leading factor in the global spread of many vector-borne infectious diseases. For example, epidemics of dengue, their seasonality, and oscillations over time, are reflected by the epidemiology of dengue in travelers (Wilder-Smith, 2008). This was exemplified during the global movement of troops (serving as susceptible hosts) and cargo ships during WWII facilitated the dissemination of the Aedes mosquitoes, and resulted in a substantial increase in the spread of the disease in Southeast Asia (Mairuhu, 2004). In addition, the transportation of used tires has been shown responsible for spreading dengue into the U.S. from Brazil and Japan in the 1980s (Wilder-Smith, 2008). Previous cases of dengue spreading between countries through infected individuals are also well documented (Gubler, 1997).

Various studies have been conducted to identify the highest travel risks with respect to vector-borne diseases, mostly in the form of surveys. One survey conducted by the European Network on Imported Infectious Disease Surveillance program (TropNetEurope, 2010), analyzed 294 patients with DF for epidemiological information and clinical features. They found most infections were imported from Asia, which suggests a high risk of DF for travelers to that region (Jelinek, 2002).

A more methodological approach was conducted by Tatem and colleagues (Tatem, 2006; Tatem, 2007). The focus of his research is on estimating the relative risk of the importation and establishment of climatically sensitive organisms (i.e. *Ae. Albopictus*) by sea and air routes. This is accomplished by remapping the global transport network to account for climate similarity. The highest risk travel routes are identified based on a normalized measure of traffic and climatic similarity. This is accomplished by first, superimposing the location (and surrounding area) of the major airports/shipping ports were onto nine gridded global climate surfaces, representing the minimum, maximum and average measurement of three different climatological variables (temperature, rainfall and humidity), defining the climate "signature" of each port. The "climatic dissimilarity" between any two ports was calculated using the Euclidean distance (of their respective climate signatures) (CED_{ij}). The total volume of travel was determined by the total number of ship visits for sea travel and total passenger volume for air travel. The product of the total travel volume between two ports and the inverse of the CED represented the link weights for the "remapped" network. The higher link weights represent a pair of well connected airports located in similar climate conditions, which increases the probability of a climactically sensitive organism relocating successfully. These weights are used to assess route risk. A separate network is created for shipping and air travel. The model also accounts for seasonality, as the monthly climate and travel volumes are disaggregate.

In an earlier paper Tatem (2006) focused specifically on the role of air travel and sea borne trade in the global dispersal of *Ae. Albopictus*, a competent mosquito vector of 22 arboviruses. Results suggest a strong positive correlation between the historic spread of *Ae. albopictus* (into previously un-established regions) and a high volume of shipping (routed from ports where the species was already established).

Tatem's work served as the main motivation for Problem III; where the focus of the proposed research is to use infection data to develop a calibrated model to assess the risk of infected vector importation and establishment based on air travel routes.

2.3.2.1 Problem III: Predicting the role of air travel in spreading vector-borne diseases

The hypothesis of this problem is that an increasing volume of international passenger air traffic originating from regions with endemic dengue increases the threat of infected vector importation, and is likely responsible for the increasing number of dengue diagnoses in the U.S. and Europe. This analysis attempts to identify those passenger air travel routes with a high likelihood for spreading infection into the United States and Europe from dengue-endemic regions. A network-level regression model is proposed which uses air traffic volumes, travel distances, predictive species distribution models, and infection data to quantify the likelihood of importing infection, relative to other routes. Thus, this research has two goals:

i. To develop a model that allows planning authorities to quantify the risk from specific air travel routes, and help identify locations where local and regional surveillance systems should optimally be implemented.

ii. To highlight the importance of proper data collection efforts that should be undertaken to enhance the predictive accuracy of such models.

There are a variety of climate sensitive biological spreading agents that are responsible for introducing diseases into previously unexposed regions of the world. For the application presented in this work the disease modeled is dengue, and the chosen vector is the *Aedes* mosquitoes (the principle spreading agent of dengue). Dengue fever is chosen for two reasons: 1) Dengue is increasingly prevalent worldwide, specifically the number of travel acquired cases reported, (today dengue is more prevalent than malaria among travelers returning to the United States from the Caribbean, South America, South

Asia and Southeast Asia (Freedman, 2006)), and 2) There is a lack of mathematical modeling tools for predicting the spread of dengue (at both the regional and global levels).

If provided with the necessary data, the model developed can be used as a prediction tool for assessing the risk of importing dengue-infected vectors or humans via air travel based on origin-destination pairs as well as to analyze the effects of changes in passenger travel routes and/or volumes on infection spreading patterns. The purpose of this research problem is to introduce the methodology and provide a sample set of results generated using existing recent data. Similar modeling tools can also be applied to any climate sensitive biological organism that could potentially be imported into a region via air travel. Additionally, the models have the potential to be extended to other transportation systems such as (rail) freight, shipping, (air) cargo, etc., other geographical regions, vector-borne diseases, other network-based processes, and even multi-layered network systems representing multiple modes of transportation in one integrated framework. Problem details are provided in chapter 5.

2.4 INTERDEPENDENT NETWORK ANALYSIS

The dispersal of infection within and across human populations is dependent on multiple network systems (e.g. social networks; transportation networks). Each of the three problems introduced focus on a single at risk network system which each plays a role in furthering the spread of infectious disease (social, transportation). These problems need to be integrated into a multi-network framework in order to develop a comprehensive epidemic prediction model. A selection of specific integration efforts was already reviewed, which introduced the current modeling techniques for integrating air travel patterns with social networks to model expected global level outbreak characteristics for contact diseases. However, these methods are specific to the S-I-R category of diseases, and are not applicable to model vector-borne disease dispersal. A simple integration of a transportation network and geographic spatial network is proposed in Chapter 5, in attempts to model climate sensitive vector dispersal on a global level. This problem is further expanded upon in the conclusions. In the following section some of the more fundamental contributions to integrated network analysis are introduced; and their potential applicability to disease prediction models is discussed.

2.4.1 Communities

Much of the literature on complex networks focuses on defining and quantifying statistical characteristics about the network topology, including centrality, density, size, motifs, hierarchical structure, and clustering (Albert, 2002); where the focus is identifying community structures at the node level ((Bagrow , 2005; Milo, 2002; Vazquez, 2004; Saramäki, 2007; Ravasz, 2002; Bianconi, 2008)). These community structures represent connections between individuals in a social network, and provide insight into the speed and direction for which a disease (or information, etc.) might disperse among the network. Similar analysis is useful for predicting disease dispersal within a social network. However, question remains: What affect does an additional network system (e.g. the introduction of additional links) have on disease dispersal within a population? For example: How does a transportation network impact the disease spreading behavior within and across regions? In order to answer this question a multilevel network system must be developed and statistical network structure properties need to be defined. A significant portion of future research effort will go towards developing a methodology for network integration processes; quantifying structural characteristics of the integrated network systems relative to the individual network systems'.

A step in this direction is taken by Palla *et.al*. (2005), who introduces an approach to analyze the statistical features of the interwoven sets of overlapping communities in efforts to identify the modular structure of complex systems. Most real world networks consist of highly overlapping cohesive groups of nodes. This work is the first to extract the traditional statistical characteristics of a complex network at the community level. In their research a community has a very specific definition; a fully connected network, which may vary in size based on an *a priori* defined number of connections each node must have. For a given complex network (with known topology) the authors employ an algorithm to identify each (overlapping) community (as defined previously), then extract information on the interconnectedness of the communities. Two communities are connected if a node in one community is linked to a node in another community. Two communities are considered adjacent if they have some predefined minimum number of connections. Based on the interconnectedness of the communities, the authors compute statistical network structure characteristics (community degree, size, clustering, etc.) at the community level. For the networks explored, their findings suggest community level statistics such as degree and clustering, share similar properties to those at the node level in complex networks.

When each community is thought of as an independent system this research provides a fundamental contribution to identifying the interaction between separate network systems. However, the form of the communities defined in this work (e.g. fully connected networks), are not representative of real world networks structures; suggesting a potential extension that should be explored. In addition, this work assumes the entire network structure (each community and inter-community connections) is known *a priori*. In regards to the analysis of interdependent systems, defining the relationship (set of connections) between these is an open problem. Intuitively such an assignment would vary with the application. For the research proposed here the application is epidemiology and the networks to be integrated are social, transportation and spatial. The appropriate methods for integration remain an ongoing research problem.

2.4.2 Infrastructure Systems

Another network integration methodology is proposed by Osorio (2005), and focuses on coupling independent infrastructure systems. This work seeks to identify the resiliency of interdependent infrastructure systems (Electric power, potable water, natural gas, telecommunications, and transportation) to internal or external disruptions (e.g., deliberate attacks, malfunction due to aging, or lack of maintenance). The interdependency among network elements is simply based on geographic proximity. The degree of coupling is defined by a tunable parameter which determines the conditional probability of one systems effect on the other (i.e. a conditional probability defines the dependence of a water pump on the power generating station it is assigned to). This parameter varies the networks from independent systems to completely dependent. They first characterize the topological properties of two interdependent small-sized real networks (representing a water distribution system and electric power network), and evaluate them when subjected to external or internal disruptions. Intuitively, they find network detrimental responses are observed to be larger when the networks are highly dependent. They conclude effective mitigation actions could take advantage of the same network interconnectedness that facilitates cascading failures.

In an additional analysis, Osorio extends the network to include three systems: gas, water and power. A network model is proposed to capture essential features of growth and evolution for interdependencies between a gas, water and power network. Dynamic response is investigated through time-dependent properties such as network resilience and fragmentation modes.

2.4.3 Human mobility networks

A different approach on integrating networks was taken by Brockmann (2006), and later extended upon (2008, 2009). This work aims to characterize human mobility patterns which encompasses both intermediate spatial scales (daily travel, car trips, etc.) and geographic global scales (air transport). To appropriately model disease spreading within human populations it is necessary to model both types of travel; however such a statistically reliable estimate of human dispersal comprising all spatial scales does not exist. The most advanced disease prediction models attempt this by coupling social networks (representing daily travel patterns) with global transport networks. One example by Balcan *et. al.* (2009), which attempts to integrate an intermediate level of transportation, regional commuting was previously introduced. Collecting transportation data for all means of human transportation is a daunting, if not impossible task, Brockmann and colleagues attempt to infer the statistical properties of human travel (on all scales, theoretically representing multiple modes of travel) by analyzing the geographic circulation of individual bank notes for which comprehensive datasets are collected at the online bill-tracking website www.wheresgeorge.com. They then identify distributions representative of “multi-scale” human movement to incorporate into dynamic disease models, among other applications. The analysis shows that the distribution of travelling distances decays as a power law, indicating that the movement of bank notes is reminiscent of superdiffusive, scale free random walks known as L`evy flights. The authors also derive a temporal aspect based on how long bills remain in a location. This is useful for developing statistical measures/distributions to sample from in large-scale simulations accounting for fluxuations between regions. A similar study was conducted by Gonzalez *et al.* (2008) to identify the spatial and temporal distribution of human mobility patterns by studying the trajectory of 100,000 anonymized mobile phone users over a six month period. This information was also used to trace the potential dispersal characteristics of mobile phone viruses which could be introduced in a cellular network.

This innovative approach at mathematically defining human mobility will serve valuable to various modeling applications. While the distributions derived to represent spatial and temporal human mobility patterns serve as an obvious asset for the different probabilistic models reviewed previously, they can also be applicable to the research problems in this dissertation to predict the impact of human travel patterns on disease spreading.

CHAPTER 3: INFERRING INFECTION SPREADING LINKS IN A SOCIAL-CONTACT NETWORK

Many factors contribute to the spread (and control) of a disease within a region, such as the standard of living, infection prevention practices (*i.e.* vaccination), local public health and emergency response programs, and perhaps most significant, the interaction patterns among individuals. Today a large proportion of the population lives in increasingly dense conditions, (e.g. modern metropolitan regions), an ideal environment for rapid disease transmission.

Significant research efforts have focused on predicting the expected spreading behavior of contact-based infectious diseases, which exploit characteristics of the population and the disease. This research compliments these probabilistic models by proposing a methodology which exploits real-time infection data to infer the most likely infection spreading scenario among a population of individuals. The types of diseases modeled in this chapter are contact-based. Contact based diseases refer to the family of infectious diseases which are transmitted from an infected to susceptible individual via direct contact, including among others sexually transmitted diseases, various strands of the flu, SARS and the common cold. Contact based diseases do not include those diseases which spread via a third spreading agent (*i.e.* a mosquito transmitting malaria). The vast majority of contact-based diseases disperse through a population in a stochastic process. For such diseases, contact between an infectious and susceptible person does not always result in a new infection. Additionally the probability of infection varies based on the disease. The stochastic nature of the infection process makes it difficult to identify the path of infection and difficult to predict the impact exposure to a new disease would have on a community, city or region.

3.1 Problem Definition

The objective proposed here is to identify the most likely path of infection (for a specific outbreak scenario) through a social contact network, by invoking the use of network based optimization algorithms and real-time infection reports (who was infected when). This research provides an alternative method for evaluating a region which has been exposed to infection; improve prediction capabilities on future potential epidemic outbreak patterns, and aid in evaluation of potential intervention strategies; without the use of computationally intensive simulations. Additionally, the solution methodology proposed has the potential to be extended to a macroscopic level model, which will be explored in the following chapter.

Similar to the model introduced by Jombart, *et al.* (2009), the methodology introduced here implements the use of a maximum probability spanning tree to capture the spatiotemporal dynamics of the infection. In this problem available infection reports, the contact network structure and disease characteristics are used to identify the most likely path of infection between individuals. The proposed model differs from Jombart's as follows: The proposed methodology is intended to 1) be implemented on a social network, incorporating network structure and related properties and 2) the edge weights are a function of the contact transmission probabilities and temporal infection data, in contrast to genetic sequencing data.

3.2 SOLUTION METHODOLOGY

The methodology described in this section uses available infection data and social contact patterns to make inferences about infection spreading patterns in a population. The network $G \in (N, A)$ can be formally defined by a set of nodes, N, which represent individuals, and links, A, which represent contacts between individuals. The problem can be further broken down into two information-based cases:

i. Full information: The ***complete set*** of infected nodes, $I \in N$, and timestamp for each infected node $t_{i,}$ (time infection occurred) is available;
ii. Partial information: Information on ***a subset*** of the infected node set, $E \in I,$ is available. This problem will be explored in future research.

Although the full information case is unrealistic with the infection data currently available, it provides a useful and necessary starting point for the proposed solution methodology, and is the case studied for the remainder of this chapter.

3.2.1 Assumptions

In order to solve the proposed problem multiple simplifying assumptions are necessary. The assumptions in this work include the following:

1. *a priori* knowledge of the underlying social contact network, $G \in (N, A)$
2. Known transmission probabilities, p_{ij}.
3. Temporal infection data is available for the full set of infected nodes, t_i $\forall i \in I.$
4. An individual can be infected at most once, thus only those diseases for which immunity is acquired after recovery are considered.

5. The outbreak evolved from a single source.

For assumption (1) the increase of social networking available on the WWW, and improvements in activity based travel modeling both contribute towards the accessibility of detailed social contact information. In regards to assumption (2), many epidemiological models assume known transmission probabilities, and ongoing research is focused on accurately quantifying these parameters. Assumption (3) is likely to be less unrealistic in the future. Increasing global internet access in conjunction with increasing global disease surveillance efforts are aiding in the availability of real time infection data. Assumption (4) just restricts the set of applications, though many diseases fall into the S-I-R category. This includes those diseases for which acquiring immunity restricts an individual from being infected more than once over the entire course of an outbreak. Under this assumption the infection spreading pattern always results in a directed spanning tree, a property which is exploited in the solution methodology. Lastly the issues posed by assumption (5) can be minimized by appropriately defining network boundaries; many outbreaks can be traced back to a single source of infection.

Subject to these assumption, the most likely spreading scenario is identified as the maximum probability spanning tree (MPST), calculated by implementing Edmonds Optimum Branching Algorithm (1967) on a sub-network connecting exclusively (known) infected nodes. The directed arcs included in the final MPST represent the most likely set of infection spreading contacts branching to the set of known infected individuals. The most computationally intensive portion of Edmond's maximum branching algorithm is the search for and removal of cycles. The assumption that individuals can be infected at most once prevents the possibility of a cycle in the outbreak scenario. Therefore the implementation of the algorithm simply requires identifying the incoming link with the highest cost for each node in the infected set *I*.

3.2.2 Link Costs

Edmond's algorithm finds the directed spanning tree, *S*, in a network where the sum of the chosen link costs included in the final tree is maximized. To implement Edmond's algorithm it is necessary to define link costs, P_{ij}, *a priori*. These link costs should be representative of the probability of an infection being spread between two individuals at a specific time.

The link probability P_{ij} is defined for link (*i,j*) as a function of 1) the link-specific transmission probability, $p_{ij} \leq 1$ (a property of the disease being modeled) and 2) timestamps of adjacent nodes, t_i and t_j. The timestamp assigned to each node is the period *t* during which the node was infected, information which is available under assumption (3). To ensure feasibility (*S* must represent a possible spreading scenario) timestamps must be increasing, and bounded along all branches in a tree. For example, if *y* is the infectious period (amount of time an infected individual remains infectious), only links (*i,j*) where $t_i < t_j < (t_i + y)$, can be included in *S*. Therefore the link probabilities, P_{ij}, only need to be computed on a subset of the links, $L \in A$. As discussed previously the link probabilities should represent the probability of infection occurring between an infected and susceptible individual at a specific time. The link probability is formally defined as follows:

$$P_{ij}(p_{ij}, t_i, t_j) = (1-p_{ij})^{(\Delta t - 1)} * (p_{ij}) \qquad \forall (i,j) \in L, \forall i \in I, \forall j \in I \qquad (3\text{-}1)$$

In (*1*) p_{ij} is the transmission probability for link *(i,j)* (that is the probability infected node *i* will infect an adjacent susceptible node *j* in a single time step; these transmission probabilities are assumed to be known under assumption 2), and $\Delta t=(t_j–t_i)$, which is the time gap between when node *i* and node *j* were reportedly infected. This function accounts for the probability that infection occurred once, p_{ij}, as well as the infection delays, or the opportunities node *i* had to infected node *j* but did not, *(Δt -1)* which has an associated probability (1- p_{ij}). The maximum number of delays should be (*y -1),* where *y* is the infectious period. This expression is related to a binomial probability, however in *(1)* order must be accounted for (the first time infection occurs the stochastic process is over). (NOTE: p_{ij} is used in reference to those social networks where the transmission probability is link specific, perhaps a function of the activity shared between nodes *i* and *j*. For other generated networks ***p*** will be used to represent a homogenous link transmission probability.)

The link cost, P_{ij} will vary with transmission probability, ***p***, and time delay, Δt. Figure 3-1 represents the P_{ij} curves for four different Δt values (2,3,4, and 5) as a function of increasing homogenous transmission probability, ***p***. The curve for $\Delta(t)=1$ is not shown, but is simply a linear function, $P_{ij}=\boldsymbol{p}$. This is simply the direct probability of infecting an individual in a single time step.

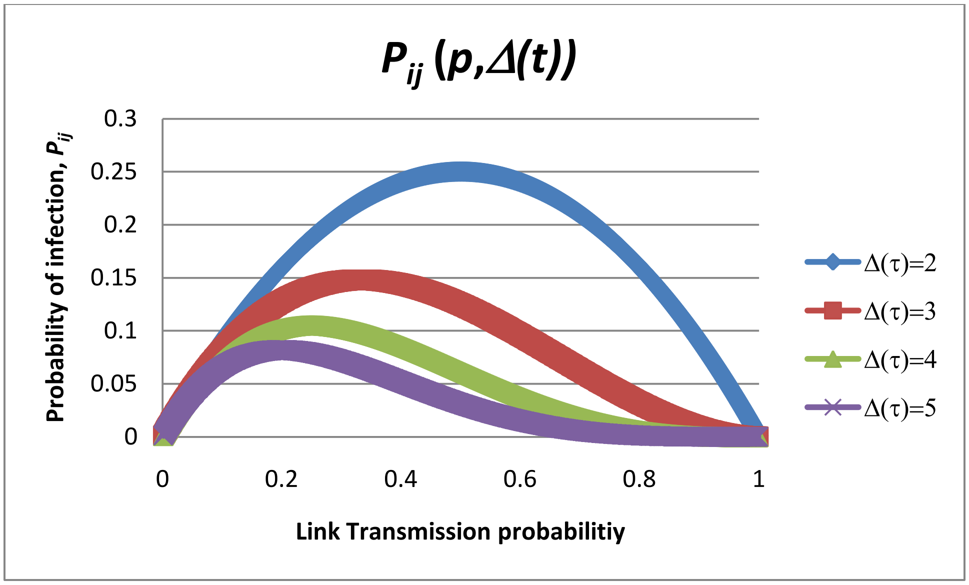


FIGURE 3-1: Probability of infection as a function of infection delay and link transmission probability

For all the curves, the probability of infection, P_{ij} decreases once a certain transmission probability is exceeded. As Δt increases, the curves become more highly skewed to the left, and heavier tailed nearing a power law function. This is because an increased transmission probability decreases the probability of a delay in the infection process (where a delay means an infected person does not infect a susceptible person they come in contact with). One important observation from this graph is the increasing difference between the *link transmission probability,* ***p****,* and the *probability of infection,* P_{ij}*,* as transmission probabilities increases. $P_{ij}(\boldsymbol{p},t_i,t_j)$ is the computed probability used in the spanning tree algorithm to predict infection causing contacts, therefore the prediction capability of the algorithm will be sensitive to the combination of link transmission probability and timestamps. For homogenous ***p*** values the link ranking will strictly depend on the Δt value. The adjacent node *i* most recently infected ($min_i \Delta t_{ij}$) will always be chosen as the predecessor node because $(1\text{-}\boldsymbol{p})^n(\boldsymbol{p}) > (1\text{-}\boldsymbol{p})^m(\boldsymbol{p})\ \forall\, n < m,$ and $0<\boldsymbol{p}<1,$ *where n* and *m* are integer values and ***p*** is the transmission probability. This also means for a given set of timestamps and homogenous ***p*** value, the algorithm will predict the same spreading scenario, as long as $0<\boldsymbol{p}<1$. While realistic social contacts networks do not have homogenous transmission probabilities, this issue is addressed because of the use of some homogenous test networks in the numerical analysis section, and the potential use in alternative applications which mind share this property.

3.2.3 Mathematical Formulation

In addition to the link cost defined above, two additional constraints must be added to the algorithm to enforce feasibility of *S*.

i. $t_i < t_j$: For all links (i,j) included in *S*, the timestamps of *j* must be greater than *i*. This enforces that for an infection causing contact between *i* and *j*, individual *i* must have been infected first.

ii. $(t_i\text{-}\ t_j) \leq y$: The timestamps for nodes *i* and *j* for any links (i,j) included in *S* must be separated by at most the infectious period, *y*. This ensures that infection causing contacts can only occur while an infected person *i* is infectious (after they are infected, before they are recovered).

Even with these additional constraints, a feasible solution, ***S*,** can always be found because these restrictions are consistent with the infection process. *S* is then computed by implementing Edmond's algorithm on a sub-network including the full set of infected nodes, *I*, the feasible set of adjacent links, *L*, with associated link costs $P_{ij}(p_{ij},t_i,t_j)$ as defined above, and the two additional time-based constraints represented as constraint (3). The formal problem definition is below:

$$max \sum_{\forall (i,j) \in S} P_{ij}\, x_{ij} \qquad (3\text{-}2)$$

s.t.

$$P_{ij} = (1-p_{ij})^{(\Delta t -1)} *(p_{ij}) \quad \forall (i,j) \in S \qquad (3\text{-}3)$$

$$t_i < t_j < (t_i + y) \quad \forall (i,j) \in S \qquad (3\text{-}4)$$

$$0 \leq p_{ij} \leq 1 \quad \forall i \in I, \forall j \in I \qquad (3\text{-}5)$$

$$\sum_{\forall (i,j) \in S} x_{ij} = |I| - 1 \qquad (3\text{-}6)$$

$$\sum_{\forall i \in I} x_{ij} = 1 \quad \forall j \in I \qquad (3\text{-}7)$$

$$x_{ij} = \{0,1\} \quad \forall (i.j) \in S \qquad (3\text{-}8)$$

$$x_{ij} = \begin{Bmatrix} 1\ if\ edge\ (i.j)\ is\ in\ S \\ 0\ therwise \end{Bmatrix}$$

The objective enforces that the set of links chosen for the spanning tree maximizes the total probability of the tree. Constraints 3-3 to 3-5 pertain to the properties and dynamics of the infection process, while constraints 6-8 enforce the spanning tree structure. Constraint (3-3) defines the link costs. Constraint (3-4) enforces that a node *i* can only infect node *j* if *i* is infected first, and still infectious (dependent on the infectious period). Constraint (3-5) restricts the link transmission probabilities, p_{ij} to be fractional. Constraint (3-6 to 3-8) together enforces that the final output is a spanning tree structure by (3-6) requiring a total of |*I*|-*1* links in *S*, where |*I*| is the number of infected nodes, (3-

7) every infected node must have one incoming link, and (3-8) the decision variable, x_{ij}, is binary.

As noted previously this problem can be solved efficiently. The maximum probability spanning tree is identified using the following algorithm:

1. Define the set of feasible links, *L*: (*i,j*) where $t_i < t_j < (t_i + y)$,
2. Calculate link costs, P_{ij} for links *(i,j)* in feasible set *L* using equation (3-1).
3. For each infected node, $j \epsilon I$, select the incoming link (*i,j*) with the highest cost, P_{ij}, from the set of feasible adjacent links, $A|j|$

The resulting tree is hereby referred to as *S*. When full information is available *S* connects *all* infected nodes in a network.

3.3 Network Structures

The networks used in this analysis are intended to represent social contact networks. For the network $G \in (N, A)$, each link (*i,j*) ϵA has an associated probability, $P_{ij}(\cdot)$ defined as in section 3.2.2.

In a contact network the links may be homogenous (all have the same probability of transmitting infection, ***p***), or heterogeneous in which case each link is assigned its own properties, p_{ij}. The properties of a heterogeneous link (*i,j*) is representative of the interaction between the two individuals (*i* and *j*) the link connects (school, work, social, etc). For randomly generated networks the structure of the network (number of contacts per individual) is determined by the degree distribution; for networks generated using actual demographic and behavioral data the network structure is a function of human activity patterns.

The proposed methodology is likely to perform differently depending on the network structure and properties of the disease. Therefore sensitivity analysis is conducted to compare the performance across various combinations of network structures (urban, power law, uniform), network sizes (in terms of number of nodes), and disease parameters (transmission probabilities). Each of the network structures generated and analyzed is described in detail below.

3.3.1 Urban Network

One of the first networks evaluated is intended to represent a social contact network for a local community of individuals that interact on a daily basis though activities such as school and work. Therefore this sample urban network will have a heterogeneous set of links, with activity-based transmission probabilities. Defining the network topology is approached with the goal of tying the social network to regional travel patterns. By using regional travel patterns (such as origin-destination tables and activity based travel patterns), individuals' daily trips and specific types of interaction (and contacts) can be accounted for. (For this model the connections are not based on any actual activity based travel data, and instead created from a synthetic data set). Unlike homogenous contact networks, the urban network designed has multiple link types, dependent on the type of trip-based contact (school, work, etc), in addition to random social links created to account for daily interactions that are not part of the traditional daily travel routine (such as contacts between family, friends, etc). As an example, a node (representing child A) might have two school links (connecting child A to child B and child C) representing contacts at school, and a social link representing contact with a neighbor they interact with after school. The set of nodes and the complete set of link types constitute the network structure. Additionally each link type has an associated probability of transmission. By discriminating between the different link types, the probabilities of transmission for different types of encounters (i.e. classmates vs. siblings) may vary, and more importantly, various real-time intervention strategies can be implemented and evaluated (i.e. closing "schools").

The urban network structure generated for this research is based on the demographic characteristics of Travis County taken from 2008 Census, such as age distribution 25%:(0-18); 65%:(18-65); and 10%: 65+, and average household size = 2.5. The network size is however a small fraction of the population (Number of Nodes) = 250. This size was chosen because it is large enough to experience complex disease spreading behavior, while small enough to evaluate for various case studies in a reasonably timely manner. The number of links in the network varies based on the level of connectivity specified when generating the network. To generate the network the following steps were taken:

1) Assign all individuals to households either of size 1, 2, 3, or 4
2) Assign all kids (individuals under 18) to a school
3) Assign all adults (individuals 18-65) to a place of work
4) Create Links connections:
    a. Connect all individuals who share a HH with Probability 1 (If two individuals share a home link, then they don't share any other links)
    b. Connect all children at the same school with each other with a Probability 0.2
    c. Connect all adults assigned to the same work office with Probability 0.1
    d. Create to random shopping links between any two nodes with Probability 0.01

   e. Create random social connections between any two nodes with Probability 0.005

5) Assign link Probabilities *p*( ), dependent on link type

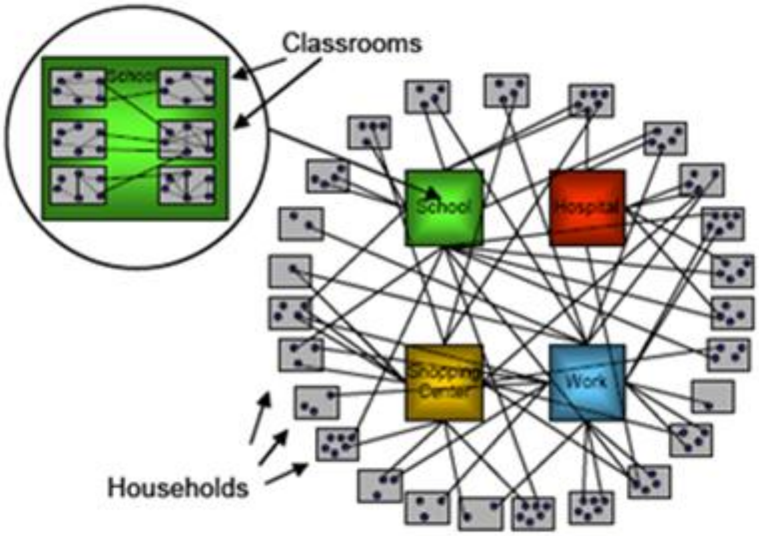


FIGURE 3-2: Example of urban network structure, Meyers *et. al.* (2005)

Figure 3-2 provides an illustrative representation of the urban network structure generated (although for the network created there are no hospitals). For the network size specified there is only one school, and five separate places of work. *p*( ) is the probability of transmission for a specific link type. This is the probability an infected node will transmit the disease to a susceptible adjacent node (which they are connected to by a specific activity) in one time step. Due to a lack of data on similar network structures and simulation models, the probabilities used to connect the network, and the transmission probabilities are chosen based on some values used in previous work by Meyers (2005) combined with my own judgment. The number of links, transmission probabilities *p*( ) and link connectivity probabilities used for the base case are listed in Table 3-1. These parameters are not based on any empirical data, but are consistent with related literature, and are purely intended to serve as a base case for analysis. Sensitivity studies were conducted to examine the level of robustness to these parameters.

TABLE 3-1: Urban network parameters for base case

Number of Links	1358
Probabilities used to create Random network Links	
Work Link	0.1
Shopping Link	0.01
Social Link	0.005
School Link	0.2
Probabilities of Transmission used in Simulation	
Home	0.2
Work	0.1
Shopping	0.05
Social	0.1
School	0.1

The probabilities used to create the links result each child being connected to about 12 other children on average. This seemed reasonable as it might be the number of students in a class, or the number of students a child interacts with each day at school. The work links results in adults being connected to 4 or 5 other co-workers on average. The social links result in a small number of close social connections, 1-2 on average, while the number of shopping links are twice that based on the assumption that an individual runs into more random people while out during the day, though with a more brief contact period. Again each individual is connected with a probability of 1 to every individual in their home, with an average household size of 2.5. The transmission probability is lowest for shopping links, $p(shop) = 0.05$. For school, work and social the transmission probability, *p*( ), is 0.1, which means that transmission occurs between these individuals on average 10% of the time a contact is made. The probability of transmission at home is the highest, at 20%. Again these are just the based case values. Sensitivity analysis is conducted by inflating and deflating these values.

Figures 3-3 and 3-4 below represent the PDF and CDF of the degree distribution for the urban network structure created. The network degree distribution is most similar to a Poisson distribution, with an average degree around nine.

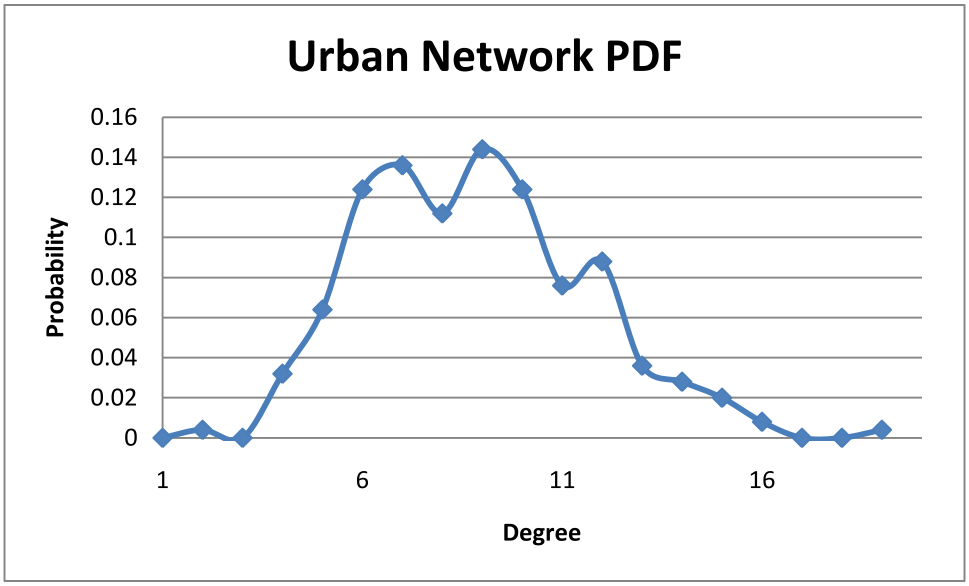


Figure 3-3: PDF of generated urban network

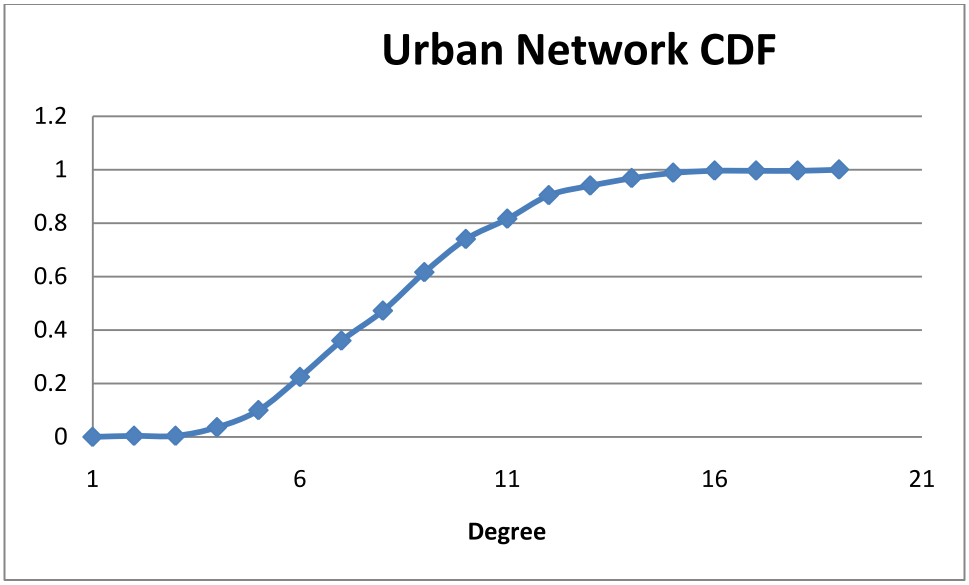


FIGURE 3-4: CDF of generated urban network

The urban network created is a relatively homogenous network structure because the majority of links have close to the average degree. To contrast this type of network, the algorithm performance will be compared with various power law network structures. A significant difference is that the power law networks generated will have homogenous link transmission probabilities, rather than activity-based transmission probabilities. However analysis will also be conducted on the urban network structure shown here, subject to homogenous transmission probabilities as well.

3.3.2 Power Law Networks

The most common network structures used for social contact network analysis are power law, which are therefore the next set of networks generated for the proposed analysis. Power law networks have a degree distribution $f(x)=ax^k$. In this analysis the exponent parameter, *k* ranges between (1, 3). The networks are generated according to

(Viger, 2005). For the power law networks generated the transmission probabilities are the same for all links, and analyzed across a range of values, (0.01, 0.5). Three examples of power law network degree distributions are shown below. The distribution with *k*=1.8 is characteristic of the global air traffic network (Kaluza, 2010), although other research has reported this value at *k*=1.5 as well. Figures 3-5 and 3-6 represent the PDF and CDF of the degree distribution for the various power law network structures generated.

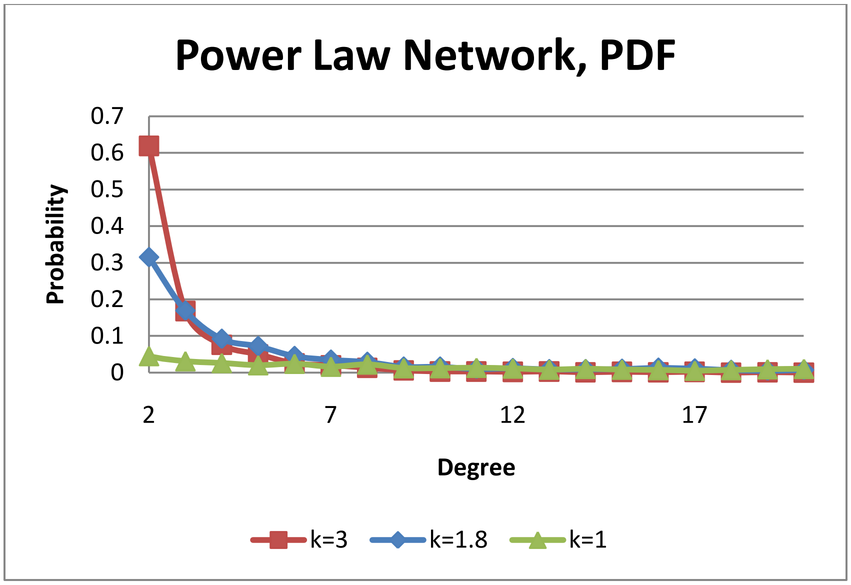


FIGURE 3-5: PDF of generated power law network

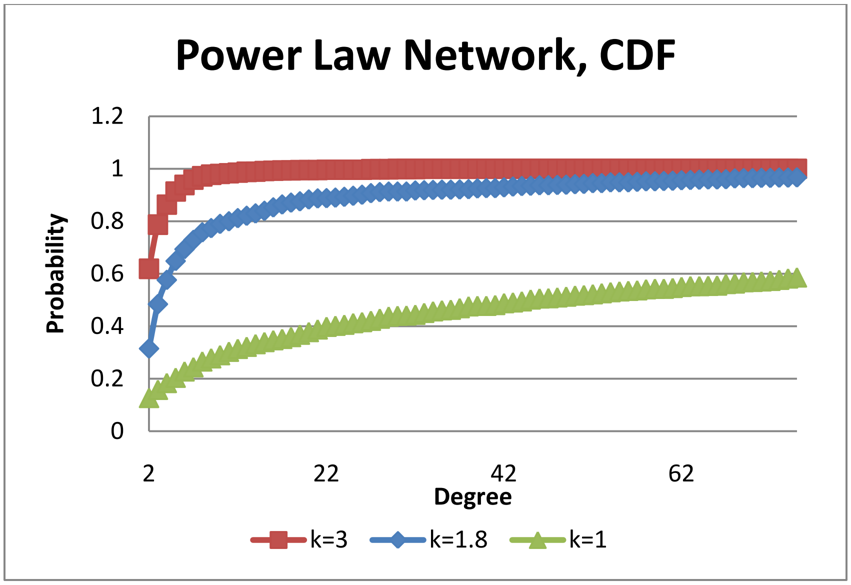


FIGURE 3-6: CDF of generated power law network

The higher exponent *k* corresponds to a more heterogeneous degree distribution. As *k* approaches one, the network structure begins to display more uniform degree characteristics. In addition, the number of links increases significantly as *k* decreases, for the same number of nodes. This is because (for the same number of nodes) there are more nodes with more connections.

3.4 Stochastic Simulation

For the case studied in this chapter the full set of input information is required to implement the proposed solution methodology. This is not currently available for an ongoing outbreak, so a stochastic simulation is used to generate input for our model, specifically the set of infected nodes *I*, and corresponding infection data, t_i (e.g. timestamps). (This research is also intended to serve as motivation for developing a ubiquitous disease database that is updated frequently, and accurately so that models such as the one proposed may be implemented for improved disease prediction and intervention strategies). Each simulation scenario defines a spanning tree where the nodes included in the tree represent the infected individuals, and the links represent the actual set of infection spreading contacts. The objective of the solution methodology proposed is to replicate this tree (the actual infection spreading pattern) as closely as possible.

Simulation Process

In this simulation model all nodes are initialized to a susceptible state, and a single node is randomly infected at the beginning of the simulation. Transmission of the disease is then simulated over multiple time steps, *t,* for a predetermined simulation period, *T*. In this model t is equivalent to the amount of time between when an individual contracts the disease and becomes infectious. This definition can vary, and at this point the *t* can be thought of in units of days. In each time step an infected individual, *i*, transmits the disease to any adjacent node *j* (that is in a susceptible state) with some known probability p_{ij}, a function of the link. If transmission occurs then the newly infected node status is changed to "infected" in the following time step, and remains so for "*y*" time steps, where "*y*" is the infectious period for a particular disease; the number of days a patient is infectious. This could also be the amount of time before recovery, hospitalization or some other type of removal from the network, during which time a node may further transmit the disease to any susceptible adjacent node. In this model "*y*" is set at 3 for the base case, though the sensitivity to this value is explored. After a node is infected for "*y*" time steps their status is changed to "recovered". Once a node is labeled as recovered they can no long transmit the disease or become infected again (this is equivalent to gaining immunity or being removed from the network). The simulation is run for a pre-determined number of time steps, *T*. Depending on the transmission probabilities set, after a certain period the total percentage of the population which becomes infected appears to stabilize.

Sensitivity of Infection Process to Disease Parameters

Various sensitivity studies were conducted to determine the robustness of the simulated infection process to the parameters chosen. For low transmission probabilities the infection process was highly sensitive, specifically those below 0.1. However at higher transmission probability values, above .15, the infection propagates so quickly that the infection process was rather robust. In addition, the model is highly sensitive to level of connectivity. Intuitively if people are highly connected then the disease will propagate much faster. Extreme examples of this can be seen when shopping, social, school and work links are removed. In the future it would be beneficial to use historical data on specific disease outbreaks as a way to calibrate the parameters used in this model. However the network structures generated are intended for evaluation of the proposed methodology, and not to recommend intervention strategies for a specific region (at this point). Therefore the structural inaccuracies resulting from the connectivity assumptions are not of vital importance at this point.

3.4.1 Using Simulation to Evaluate Outbreak scenarios and intervention analysis

The objective of most epidemiological models is to develop the most effective disease control and surveillance measures, where the total number of infections is minimized. As mentioned in the literature review, computationally expensive microscopic stochastic simulation models are often used to evaluate various possible intervention strategies such as closing schools and work, minimizing social interaction, or quarantining individuals. Examples of this type of analysis are conducted and presented, which also illustrates the sensitivity of the infection process and types of parameters identifies previously.

As an example, for the urban network structure used in this work the impact of reducing contact among individuals or preventing certain types of interactions is be analyzed by removing some or all of a chosen link type from the network (thus transmission can no longer occur), or reducing the transmission probabilities for a specific family of link types. If it is decided that schools should be closed, then all links of type “school” can be removed from the network, and disease transmission will no longer occur on any of those links. Additionally, this type of decision may be implemented at any specified time step during the simulation. It is therefore important to explore and evaluate various potential intervention strategies that can be feasibly implemented, and the impact of implementing them at various times throughout the progression of the disease. The time based decision is important because it is not always feasible or economical to implement certain strategies such as school or work closure when there are very few infections reported, on the other hand, after a disease has infected some percentage of the population (sometime referred to as the epidemic threshold) certain intervention strategies may no longer be effective.

The microscopic simulation model provides information on 1) The average cumulative population infection levels over time, which represents how fast the disease is spreading through the population, and 2) The average infection level by activity type and therefore the relative transmission levels of each activity type. This information allows us to identify which types of activities (social, shopping, school, etc.) are responsible for the majority of transmission of the disease (see Figure 3-8 and 3-9). Based on the results from the simulation, the following intervention strategies were explored:

i. Vary the infectious period, "y", of the disease. This illustrates the benefit of better diagnostics (a faster diagnosis theoretically results in earlier removal of an infected individual from the network which may be possible through improved isolation measures and governmental regulations over time as disease progresses.). It also illustrates the how cumulative infection rates vary as a function of the infectious period of a given disease.

ii. Improve isolation measures by limiting random daily interaction (i.e. reduce the number of shopping and social trips made)

iii. Reducing school and work interaction by closing school and work offices. This decision strategy is implemented at each time step of the simulation so as to evaluate the effect of delaying this type of decision.

iv. Vary the probability of transmission, P( ), of the disease. This study reveals the benefit of reducing one's exposure levels by wearing masks or even washing hands. This case study also illustrates how infection rates vary as a function of how contagious a disease is.

For the sample simulation results below, *T*=10, and results are averaged over 250 iterations in order to capture the average behavior of the disease. In each iteration the initial infected node is randomly selected to account for the difference in transmission behavior when an adult is "patient zero" versus a child or an elder. The number of iterations was chosen as 250 because the average network behavior appeared to converge after about 200 iterations. This simulation captures the average number of infected nodes in the population, the link types that transmit the disease, as well as the average temporal progression of the disease. By keeping track of the link types which spread infection, the types of activities which are on average responsible for the infection are exposed, and this information can then be used to develop decision strategies to prevent further spread.

The simulation results shown in Figure 3-7, 3-8, and 3-9 are all for the case where the intervention strategy was implemented at the beginning of the simulation, t = 0. This is knowingly unrealistic because intervention would likely not be implemented until it was evident that some potential outbreak was a possibility, however simulating in this manner reveals the relative role each activity plays in transmission of the disease, and implementing each decision at t=0 reveals the relative benefit of each intervention strategies, which may therefore be implemented at any time stage of the outbreak. Figure 3-7 and 3-8 provide cumulative infection rates and number of activity based infections from five different case studies. Results from the base case, shown in dark blue, were used to develop potential intervention strategies. The resulting infection levels should be analyzed relative to one another, rather than quantitatively because each case is a variation of the base case, and the true quantitative values are dependent on the chosen parameters.

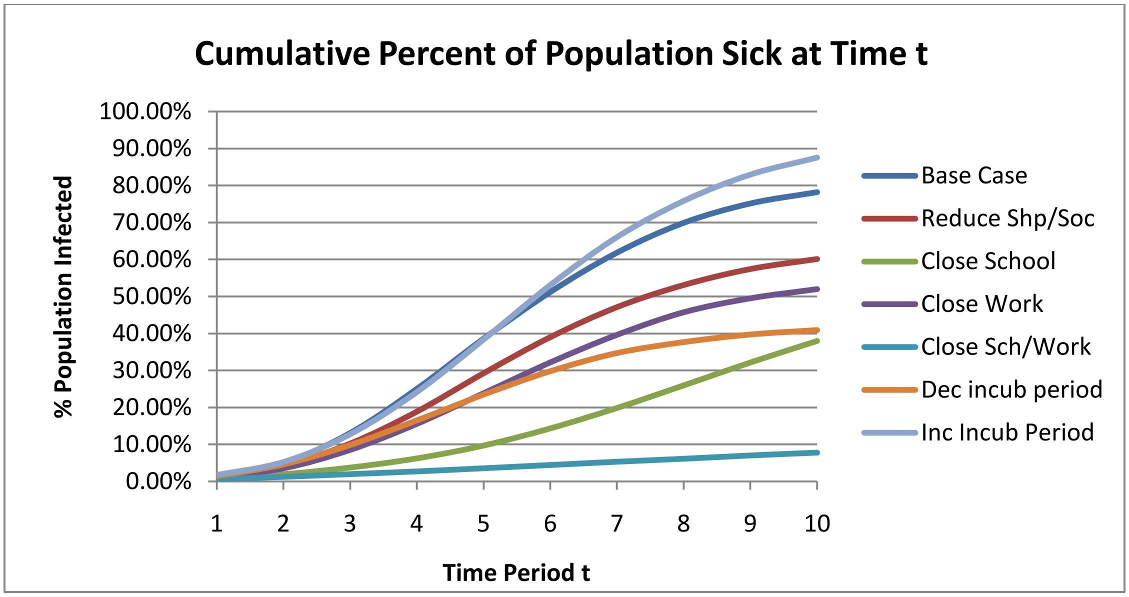


FIGURE 3-7: Cumulative percent of population sick at time t

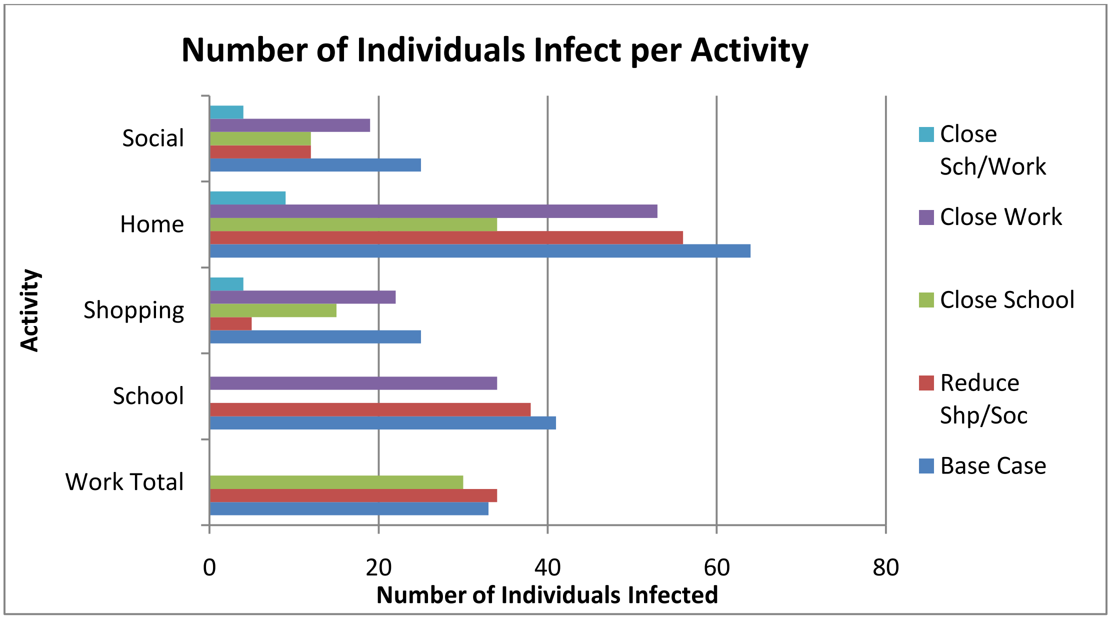


FIGURE 3-8: Number of individuals infected per activity

From Figure 3-8 it is clear that a large portion of the infections are based out of the home, likely a result of the close contact levels among family members, and therefore higher rates of transmission. Home contacts are logistically less feasible to prevent, however targeted home antiviral treatments may be a beneficial option because of the relatively high rate of transmission. The second highest number of infections is caused at school. This is expected because of the high level of interactions during a school day, and less precautious nature of children. From this result, school closure is often chosen as an intervention strategy.

Figure 3-9 provides the effect of varying the infectious period, the amount of time an infected individual remains on the network in an infectious state. Again the dark blue is the base case, with y=3. The light blue represents an increased infectious period, y=4, while the orange represents a decreased infectious period, y=2. It appears that increasing the infectious period by one does not have near as significant affect as lowering it by one. When y=2 the total number of infections is reduced by half among all activities. This implies that improved diagnostics, the ability to catch the disease early, or quarantining infected individuals could drastically reduce the total number of infections that arise. On the other hand, an increased infectious period results in a 10% increase in the cumulative infected level after 10 time steps, but does not vary much from the base case before the 6th time step.

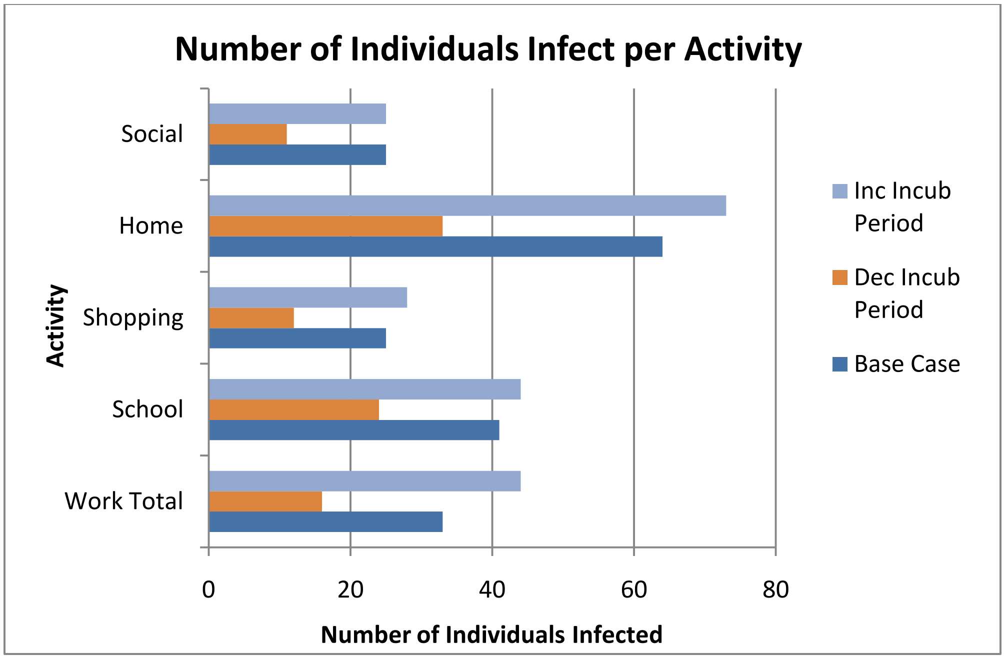


FIGURE 3-9: Number of individuals infected per activity

The previous analysis using stochastic simulation is an example of a probabilistic tool that is often used for predicting future and analyzing past outbreak scenarios, and evaluating various intervention strategies. This analysis is intuitively sensitive to the simulation and disease parameters chosen. While sensitivity analyses can be very useful in evaluating the robustness of various intervention strategies, extensive analysis can be computationally very expensive.

The methodology evaluated in the following section is still probabilistic in nature, as it is a function of a stochastic spreading process; however it uses network based optimization methods to predict the single most likely outbreak scenario by identifying the contacts (network links) which are most likely responsible for spreading infection. This information can then be inferred to come up with potential intervention strategies, similarly to the way stochastic simulation analyses are used. Similar parameter sensitivity analysis is conducted for the proposed methodology.

3.5 Measure of Performance

Although the solution method itself does not require the use of a microscopic level stochastic simulation model, one as defined in section 3.4 is used in to evaluate the performance of the proposed methodology. The algorithm requires the following data for input: the full set of infected individuals, *I,* the contact network structure $G \in (N, A),$ the time each individual was infected, t_i (to extract the set of feasible links, *L*), and link transmission probabilities, $p_{ij},$ both used to calculate the link costs, $P_{ij}(p_{ij},t_i,t_j)$ for the set of feasible links *L*. The performance is measured by comparing the set of links from the simulation-based scenario (*e.g.* the predecessor node for each infected node), with those in the model output *S*. To evaluate the performance of this solution methodology the following formal steps are taken:

1. Generate a network, $G \in (N, A),$ and specify link transmission probabilities, p_{ij}
2. Randomly introduce an infected individual into the network
3. Simulate an infection spreading scenario for some preset time period, *T*
4. Extract the following (required) information from the simulation to use as input for the solution methodology
    a. Full set of infected nodes, *I*
    b. Timestamps for each infected node, t_i $\forall i \in I$.
5. Extract the following information from the simulation to use for evaluating the solution methodology
    a. Full set of links in the infection tree, *K*
6. Implement the solution algorithm (steps 1-3 in section 3.2.3) (on the extracted network $G \in (I, L)$):
    a. Identify the feasible link set, *L*

    b. Calculate link costs, P_{ij}
    c. Compute maximum cost spanning tree, *S*.
7. Identify the percentage of correctly predicted links, *q*.
    a. Identify the set of links *M*∈*K*, which are the links in *K* that are also in the model output *S*.
    b. *q*=|*M*|/|*K*| (This is the percentage of links that are correct in *S*)
8. Repeat steps (2)-(6) X times, and average *q* (step 6) over all iterations.

This procedure returns the *expected* performance of the solution methodology, ***Q***, which is how closely *S* represents the actual spreading scenario, on average. This analysis is performed for various combinations of network structures, sizes, and disease parameters. The results are presented in the following section.

3.6 Numerical Results and Analysis

The expected performance of the solution methodology, ***Q***, calculated using steps 1-7 defined as in the previous section, is illustrated in Figure 3-10 and 3-11 for the following network structures, respectively:

i. Urban network

ii. Power law network

For both network structures ***Q*** is an average from X=1000 iterations, and presented for various combinations of *y*, and *T*. In the figures each series represents ***Q*** for a constant infectious period, *y* (*y* is shown as an integer in the legend). The results illustrate a decrease in expected performance, ***Q*,** as the simulation time, *T* increases. ***Q*** also decreases as the infectious period, *y*, increases for a constant simulation period, *T*.

This is an intuitive result, as the uncertainty increases with higher infectious periods and simulation times due to the stochastic nature of the infection spreading behavior.

The urban network used in this analysis has 250 nodes, 979 links and the original transmission probabilities as defined in Table 3-1, varying between (.05, 0.2). The power law network also has 250 nodes, though fewer links because of the heavier tailed distribution (relative to the Poisson-characterized urban network). The power law network used in this analysis has 379 links, and *k*=3.

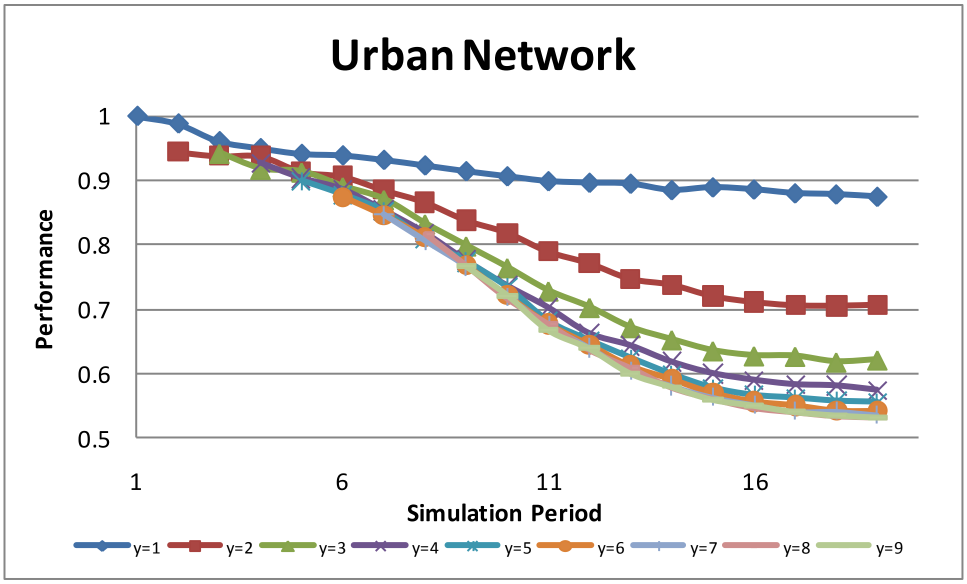


FIGURE 3-10: Expected performance, *Q,* for urban network

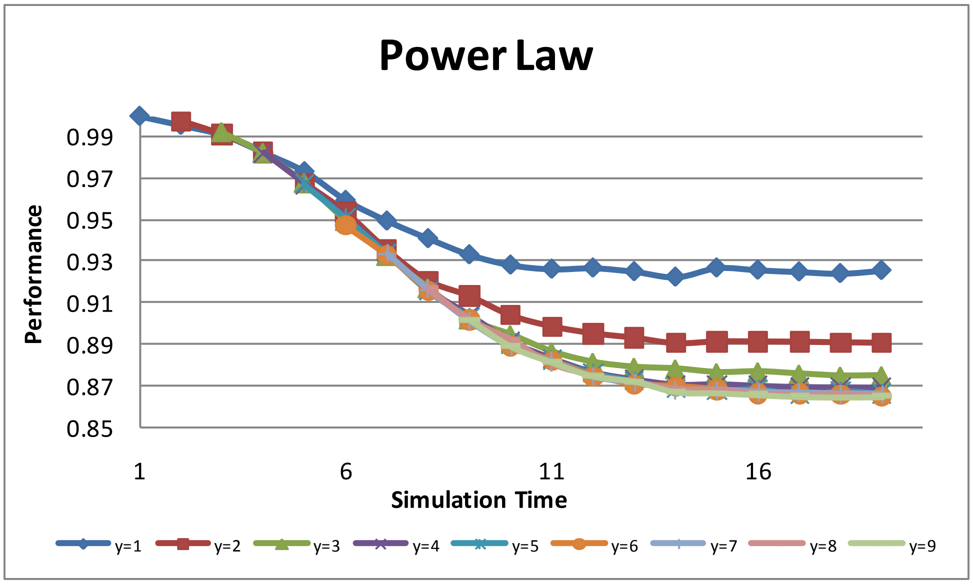


FIGURE 3-11: Expected performance, *Q*, for power law networks

From the results, it appears ***Q*** is rather robust for *y*>4, and *T*>14, or the equivalent of two weeks. From figure 3-7 it is clear that (with the chosen set of parameters) after a period of time the size of the outbreak stabilizes. Figure 3-7 also shows the increased rate at which infection spreads for higher *y* values. The stability of the infection process under these conditions is likely the explanation for the robust performance of ***Q.*** These results also suggest that diseases with reduced levels of exposure (shorter infectious periods) will have much more predictable outcomes, especially during the first week of the outbreak.

Overall the same basic behavior is present for both the urban and power law network structures, although ***Q*** performs better for the power law network remaining above 85% for all cases analyzed, while in the urban network ***Q*** falls closer to 50%-60% for higher (*y,T*) combinations. The variation in performance is likely a results of the homogenous transmission probability chose for the power law network, ***p*** = 0.5, compared to the heterogeneous urban network structure. There are fewer links in the power law network studied which is also a likely factor in the increased performance. Sensitivity analysis for the transmission probability is presented in the following section.

3.6.1 Sensitivity to Transmission Probability Value

Sensitivity analysis is conducted for urban and power law network structures for varying transmission probabilities, a constant infectious period, y=3, and simulation time, T=15. These values (y, T) were chosen because they appear to represent stable parameters in the analysis above. The results will differ for different combinations of y and T, however the purpose of this analysis is to uncover the performance trends as a function of transmission probability; the actual performance should not be assumed to be representative of all cases. The value chosen for T is rather high, so these results likely underestimate $\boldsymbol{Q}$, but still serve as an appropriate basis for comparison for sensitivity analysis.

3.6.1.1 Urban Network

$\boldsymbol{Q}$ is illustrated for the urban network subject to varying transmission levels (which increase along the x-axis) in Figure 3-12. The urban network with the original activity-based link transmission probabilities is referred to as Urban Network I. Again the results are averaged over X=1000 iterations. The original transmission probabilities are inflated and deflated by a constant factor, thus remaining proportional to the original activity-specific values. The maximum inflation factor is five, at which point some of the links take on a transmission probability value of one. The maximum deflation is 0.01, which results in many p_{ij} values close to zero. These inflation and deflation factors are chosen such that ($0<p_{ij}<1$).

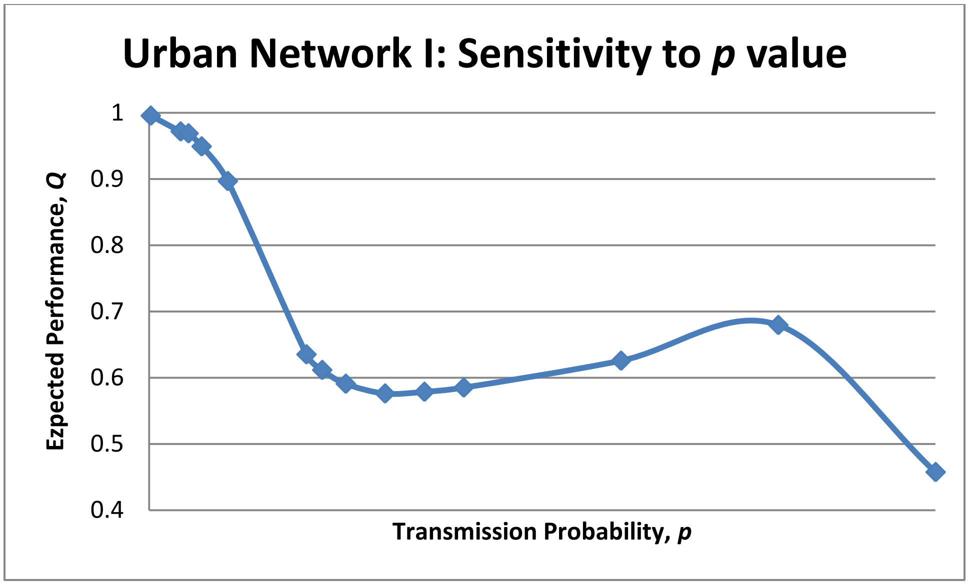


FIGURE 3-12: Heterogeneous urban network sensitivity to transmission probability

Q is remarkably close to one at low transmission probabilities due to the low level of uncertainty associated with the infection process. In this case the transmission probability is nearly zero, and the infection process is nearly deterministic. If all the transmission probabilities were zero or one, then the exact scenario would be predicted for *S* because the infection process would actually be a deterministic process, and the infection causing predecessor nodes would be known with certainty (with the exception of ties, when two possible nodes could have resulted in the infection of another). The lower performance when p_{ij} is inflated is likely a function of the extreme differences in transmission probability values across links, which range from .05 to 1.

In Figure 3-13 the urban network links are assigned homogenous transmission probabilities (as is the case in the power law network). This network is hereby referred to as Urban Network II. $\boldsymbol{Q}$ =1 for the deterministic scenarios ($\boldsymbol{p}$=0 and $\boldsymbol{p}$=1), and takes the lowest value when $\boldsymbol{p}$=0.2. As illustrated in Figure 3-13, $\boldsymbol{Q}$ varies with transmission probabilities. The increase in performance as the transmission probability increases can be explained using the known algorithmic behavior for networks with homogenous transmission probabilities: *the link with the smallest infection delay is always chosen as the infection causing contact* (see section 3.2.2). For higher $\boldsymbol{p}$ values this is more likely to be accurate (higher transmission probability will more often result in immediate infection). Similarly, the steep decrease in $\boldsymbol{Q}$ as $\boldsymbol{p}$ increases from the deterministic case, $\boldsymbol{p}$=0, can be attributed to the fact that a longer infection delay is more likely to occur at lower transmission probabilities, although the link with the smallest delay is always going to be chosen, therefore S is less likely to replicate the actual spreading scenario.

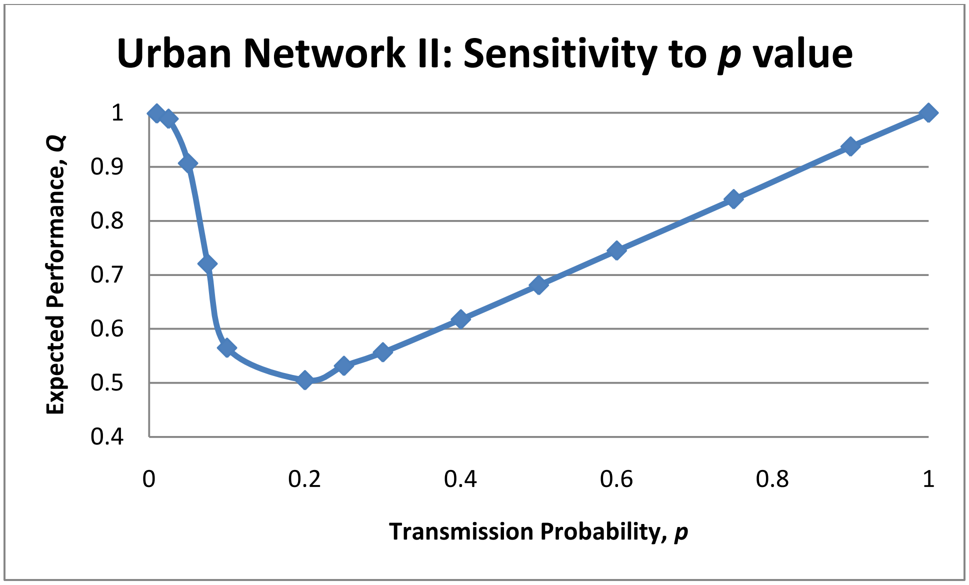


FIGURE 3-13: Homogenous urban network sensitivity to transmission probability

3.6.1.2 Power Law Networks

Similar to the analysis presented previously for the urban network, the robustness of ***Q*** to transmission probability is explored for power network structures. For the first analysis presented on a power law network (in figure 3-11) the transmission probability was set at a constant value (***p*** = 0.5) for all links, meaning there was a 50% chance of an infected individual infecting a susceptible one each time step. This is representative of a highly infectious disease. To illustrate the robustness of ***Q*** to the transmission probability, ***Q*** is presented for 14 values of ***p,*** ranging between (0,1), for each of the three power law networks. The analysis was conducted for the same constant infectious period *y*=3, and *T*=15, and results averaged over 100 iterations (which is lower due to the larger network size, and increased running time). For each of the three networks, ***Q*** is illustrated for each of the assigned link transmission probabilities in Figure 3-14.

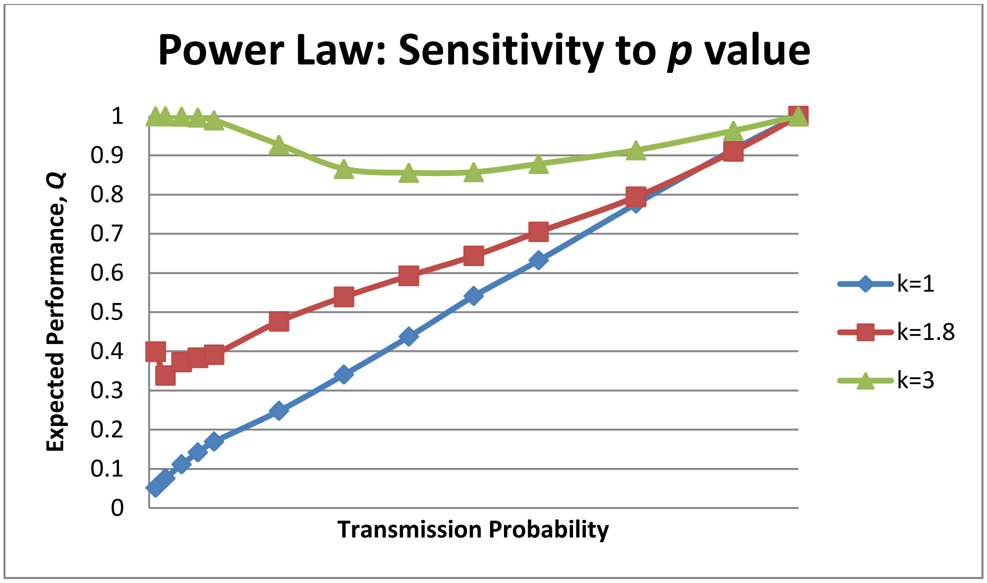


FIGURE 3-14: Power law network sensitivity to transmission probability

The specific properties of each of the power law networks and the simulation parameters are summarized in table 3-2. Each of the power law networks generated for this analysis has 1000 nodes, while the number of links varies based on *k*. The number of nodes and lower and upper bounds for node degree, along with the exponent, *k* are specified as input to generate a network with the defined properties.

TABLE 3-2: Power Law Network Properties

# Nodes	# Links	K	Lower Bound	Upper Bound	y	T
1000	1565	1	2	1000	3	15
1000	7531	1.8	2	1000	3	15
1000	81726	3	2	1000	3	15

For the power law networks, the performance varies significantly as a function of network structure and transmission probability. The most heterogeneous network structure (*k*=3) performs best for all ***p*** values, while the most homogenous (*k*=1) performs worst. It is likely that the significant increase in the number of links for the *k*=1 network contributes to the poor performance at very low transmission probabilities. The highly heterogeneous network structure likely contributes to the improved performance at low transmission levels, because the probability of infecting a hub is extremely low for a low transmission probability, which means the variability in spreading scenarios remains minimal relative to a more connected network, as is the case when *k*=1.

For all the power law network structures (though it is not shown on this graph), ***Q*** =1 for ***p***=0, representative of a deterministic case. For the most heterogeneous network structure, *k*=3, ***Q*** approaches one even for low, non-zero transmission probabilities. This contrasts the other two power law networks in which ***Q*** =1 only when the spreading scenario is fully deterministic, and quickly decreases for low ***p*** values. For all networks ***Q*** improves nearly linearly as the transmission probability increases, though ***Q*** begins increasing at a lower ***p*** value lowest for *k*=1.8 and *k*=1. The same logic as in the homogenous urban network results (Figure 3-13) can be applied here. Because the most recently infected predecessor node will always be selected in a network with homogenous ***p*** values, the predictions are more likely to correspond to cases with a higher transmission probability. This fault is exaggerated for the more homogenous network structures.

3.6.2 Sensitivity to Transmission Probability Accuracy

The next analysis presented explores the sensitivity of $\boldsymbol{Q}$ to the accuracy of the $\boldsymbol{p}$ value assigned**.** This is representative of a situation where a disease's properties are *unknown*. This analysis differs from the previous analysis which explores the methodology performance subject to various *known* $\boldsymbol{p}$ values. To explore the robustness of the model to accuracy of $\boldsymbol{p}$, a $\boldsymbol{p'}$ value is selected which differs from the actual $\boldsymbol{p}$ value by $\boldsymbol{\Delta p}$**:** $\boldsymbol{p'=p+\Delta p}$. This estimated $\boldsymbol{p'}$ is used in the link costs to calculate *S*. $\boldsymbol{\Delta p}$ can be positive or negative, as long as $0<(\boldsymbol{p+\Delta p})<1$, and simply represents the inaccuracy of the transmission probability value assumption. For example, when $\boldsymbol{p}$ = 0.5 (this is the actual transmission probability which dictates the behavior of the outbreak), and $\boldsymbol{\Delta p}$ = -0.3, *S* is determined using $\boldsymbol{p'}$ = 0.2, and not the true value, $\boldsymbol{p}$ = 0.5. This would be a case where the disease is thought to be much less contagious than it actually is. $\boldsymbol{Q}$ is still calculated in the same way, by comparing the actual spreading scenario (here represented using a simulation with a specified $\boldsymbol{p}$), with $S_{p'}$ determined using $\boldsymbol{p'}$**.** The impact of using $\boldsymbol{p'}$ (instead of $\boldsymbol{p}$) on $\boldsymbol{Q}$ is illustrated by comparing $\boldsymbol{Q_p}$ and $\boldsymbol{Q_{p'}}$**.** These values are the expected performance under the original assumption that the correct $\boldsymbol{p}$ is known when solving S_p, $\boldsymbol{Q_p}$**,** and the expected performance when $\boldsymbol{p'}$ is used to determine $S_{p'}$, $\boldsymbol{Q_{p'}}$. $\boldsymbol{\Delta Q}$ = ($\boldsymbol{Q_p}$ - $\boldsymbol{Q_{p'}}$**).** For the networks with a homogenous link transmission probability $\boldsymbol{p}$=0.5, the $\boldsymbol{\Delta p}$ varies between (-0.5, 0.5). Figure 3-15 illustrates the sensitivity of $\boldsymbol{Q}$ to the accuracy of $\boldsymbol{p}$, $\boldsymbol{\Delta Q}$**.** This analysis is conducted for the urban (blue) and power law with k=3 (red) network structures with homogenous link transmission probabilities, $\boldsymbol{p}$=0.5, y=3, and T=15.

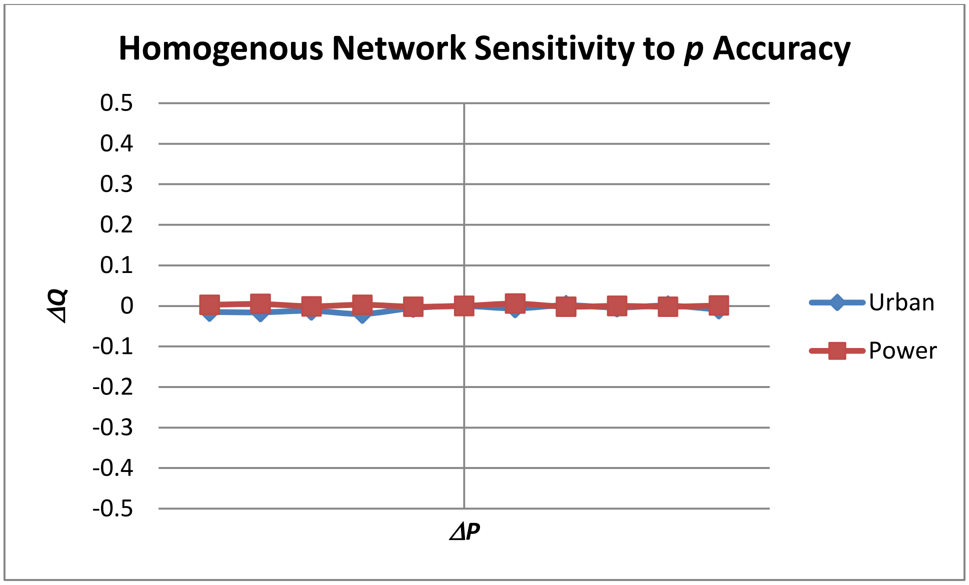


FIGURE 3-15: Homogenous network sensitivity to accuracy of transmission probability

It is obvious from the graph that $\boldsymbol{Q}$ is highly robust to varying levels of inaccuracy in the transmission probability used to infer the outbreak scenario. At the extreme left and right of the series the $\boldsymbol{\Delta p}$ is -0.5 and 0.5. For this range of error the expected performance deviates by less than 2%. This is because for a network with homogenous transmission probabilities, $\boldsymbol{p}$, and the link cost function $P_{ij}(\boldsymbol{p},t_i,t_j) = (1-\boldsymbol{p})^{(\Delta t-1)} * (\boldsymbol{p})$, an overestimation or underestimation of $\boldsymbol{p}$ by some $\boldsymbol{\Delta p}$ will affect each link cost proportionally:

$P_{ij}(\boldsymbol{p'},t_i,t_j) = (1-\boldsymbol{p'})^{(\Delta t-1)} * (\boldsymbol{p'}) = P_{ij}((\boldsymbol{p+\Delta p}),t_i,t_j) = (1-(\boldsymbol{p+\Delta p}))^{(\Delta t-1)} * (\boldsymbol{p+\Delta p})$

S is defined by selecting (from the set of feasible links) the incoming link with the highest link costs for each infected node. As long as $\boldsymbol{p}<0$ and $\boldsymbol{p'}<0$, the rank of incoming links will remain constant, therefore $S_p=S_{p'}$. Again, for homogenous $\boldsymbol{p}$ values the link ranking will strictly depend on the Δt value (*e.g.* the adjacent node i most recently infected ($min_i \Delta t_{ij}$) will always be chosen as the predecessor node). This only applies when

$\boldsymbol{\Delta p}$ is the same for all links, or all transmission probabilities are over/under valued equally. The minimal variance seen in Figure 3-15 is a function of the stochasticity of the infection process, and the number of scenarios being averaged. If X was increased, then $\boldsymbol{\Delta Q}$ would converge to zero.

If the transmission probabilities are heterogeneous S_p is not necessarily going to be the same as $S_{p'}$. Using the familiar link cost function and a constant $\boldsymbol{\Delta p}$ across all links ($p'_{ij}=p_{ij}+\boldsymbol{\Delta p}$):

$P'_{ij}(p_{ij},t_i,t_j) = (1- p'_{ij})^{(\Delta t -1)} *(p'_{ij}) = P_{ij}(p_{ij} +\boldsymbol{\Delta p},t_i,t_j) = (1-(p_{ij} +\boldsymbol{\Delta p}))^{(\Delta t -1)} *(p_{ij} +\boldsymbol{\Delta p})$

And the simple three node network (shown in Figure 3-16):

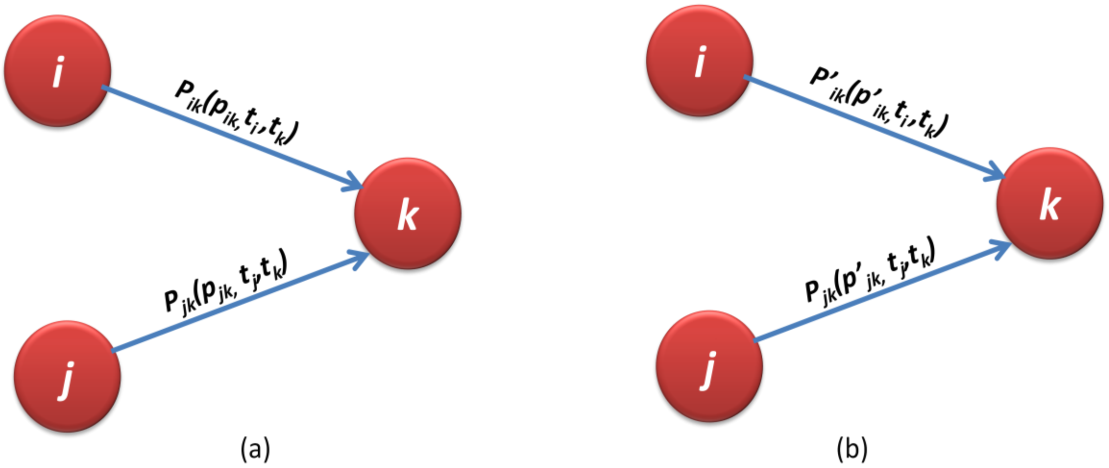

i
$P_{ik}(p_{ik}, t_i, t_k)$
k
$P_{jk}(p_{jk}, t_j, t_k)$
j
(a)
i
$P'_{ik}(p'_{ik}, t_i, t_k)$
k
$P_{jk}(p'_{jk}, t_j, t_k)$
j
(b)

FIGURE 3-16: Example network and link costs for (a) network with accurate transmission probabilities, p and (b) inaccurate transmission probabilities, p'

If $P_{ik}> P_{jk}$ it is not always true that $P'_{ik}>P'_{jk}$. For example if p_{ik}=0.2, p_{jk}=0.5, $t_i=t_j=1$, and t_k=3: $P_{ik}= (1- p_{ik})^{(\Delta t -1)} *(p_{ik})=(1-0.2)^{(3-1-1)}*(0.2)=(0.8)*(0.2)=0.16$ and $P_{jk}= (1- p_{jk})^{(\Delta t -1)} *(p_{jk})=(1-0.5)^{(3-1-1)}*(0.5)=(0.5)*(0.5)=0.25$. In this case $P_{ik}< P_{jk}$. However, when $\boldsymbol{\Delta p}$=0.3, p'_{ik}=0.5, p'_{jk}=0.8, and P'_{ik} =0.25 and P'_{jk} =0.16, in which case $P'_{ik} > P'_{jk}$, and $S_p \neq S_{p'}$. This provides a counter example to prove that the initial ranking of adjacent links will not always remain constant, though there are cases where $S_p = S_{p'}$ for heterogeneous link functions. For example when $\Delta t = 1$ for all adjacent links (of a given node), $P_{ij}(p_{ij}, t_i, t_j) = p_{ij}$, and as long as all the link transmission probabilities are inflated/deflated by a constant $\boldsymbol{\Delta p}$, $P'_{ij}(p_{ij}, t_i, t_j) = p_{ij} + \boldsymbol{\Delta p}$, and $S_p = S_{p'}$.

The difference in prediction capability is illustrated in Figure 3-17, which is analogous to the method used to create Figure 3-15, but for the urban network I, with heterogeneous link transmission probabilities. The $\boldsymbol{\Delta p}$ is chosen such that $0 < p_{ij}' < 0$ for all links. The results are still rather robust until the $\boldsymbol{\Delta p}$ increasing to the point that some of the transmission probability values approach one. At this point the expected performance reduces by nearly 25% from that which would be obtained using the actual p_{ij} values.

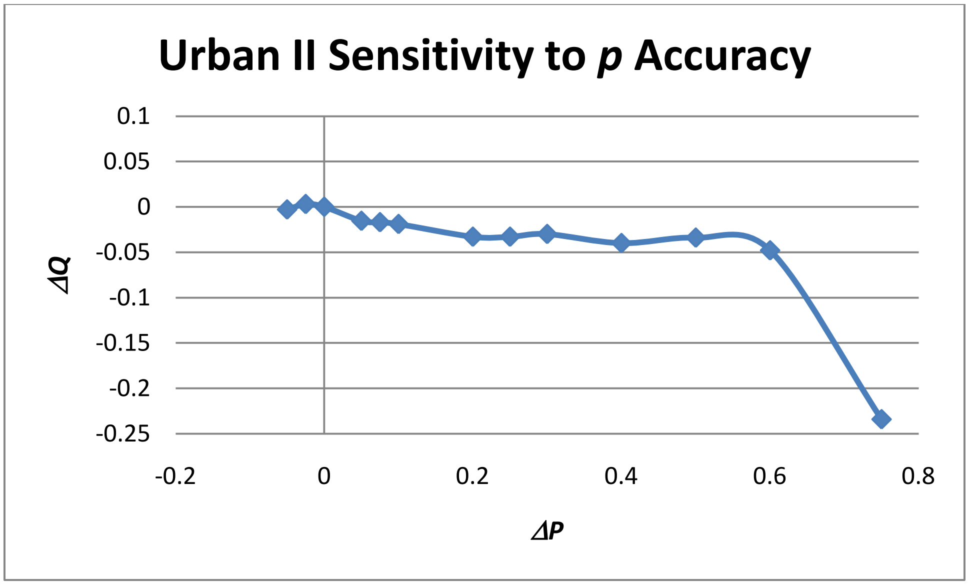


FIGURE 3-17: Heterogeneous network sensitivity to accuracy of transmission probability

3.7 Conclusions and Future Research

These analyses provide insight into the performance of the proposed methodology as a function of network structure, network size, disease properties as well as potential human error in assessing the disease properties.

While the performance varies significantly as a function of network structure and transmission probability, the methodology performs best for the most heterogeneous network structures. This is favorable because heterogeneous structural properties are characteristic of many real world networks on which infection processes occur. In addition the methodology performs best for a lower range of transmission probabilities among links in a network. This is a property which may or may not pertain to a realistic network structure subject to an infection process, dependent on the disease and definition of the links. Lastly, the performance appears to be robust to modest estimation errors in terms of transmissibility, even for networks with heterogeneous transmission probabilities. This is another favorable characteristic because accurate disease transmission properties are difficult to estimate. One potential future research plan might explore ways to transform link properties and/or the network structure itself such that the network modeled reflects the properties for which the methodology performance is maximized.

The results from this chapter prove insightful for extensions of this methodology, such as the macroscopic version of the problem introduced in the following chapter, which seeks infection spreading travel routes (links) between regions (nodes) rather than individuals. The main research focus for the macroscopic version is defining link costs functions represent of the probability of infection occurring across regions. The sensitivity of ***Q*** to variations in link costs and network structure, illustrated in this chapter, will aid in the development and analysis of link costs in the next chapter.

In addition to the macroscopic version of the problem, the proposed model has multiple potential extensions which will be expanded upon in future work. Two of which are introduced below.

Intervention Strategy Analysis

One of the potential uses for this type of model is evaluating proposed intervention strategies and policies. This type of analysis is specific to network structures such as the urban one used here, which have a heterogeneous set of links. *S* identifies the set of infection links, which can provide *insight* into the spreading behavior of a disease because the contact types most likely to spread infection are revealed. This is analogous to the simulation based analysis to extract the expected role each activity played in the disease spreading process (Figure 3-8); achievable with a single iteration of the proposed algorithm. For example, if a high percentage of links in *S* are school links, this suggests an effective intervention policy would be to temporarily close schools. Furthermore, if information was available such that individual schools could be distinguished by link type, policies can be specified at the individual school level.

An additional analysis possible within this methodological framework is to *compare* intervention strategies. The probability of a given spreading scenario, *S,* can be compared when various intervention policies are implemented (e.g. strategically removing certain link types, or adjusting the transmission probabilities). If the probability of a spreading scenario is significantly reduced after a set of links is removed, this suggests such contact restrictions might reduce the spread of disease. Additionally, a hierarchical evaluation of intervention policies is made possible; for each intervention strategy implemented, the second (and third and so on) most likely alternative spreading scenario is revealed.

For the intervention analyses discussed a level of detail on the network structure is required which may not be available (specifically link types should be differentiated). This issue reveals the inevitable tradeoff between data availability (in terms of network structure, contact specific transmission probability, information, etc. provided as input for the model), and the level of analysis that can be provided. Additionally this methodology does not explicitly provide support for temporal-based intervention strategies, (e.g. when and for how long should a school be closed?).

This problem should also motivate the development of an open source infection database for researchers and medical personnel.

Extension to the Partial Information Case

A methodological extension of the problem is defined in the chapter as case 2; relaxing the full information assumption. This includes evaluating the performance under different assumptions of available information. The first problem proposed assumes an availability of complete infection data for the set of infected individuals in a community (undefined size), and attempts to predict the contacts responsible for spreading infection. A further extension of this problem represents the more realistic setting where only a fraction of infected individuals consult a physician, visit a hospital, etc., resulting in partial information. The objective is again to determine the most likely set of infection spreading contacts resulting in a known set of infected nodes, when partial infection data is available (so only a subset of the infected nodes are identified, $E \in I$). A further complication arises when the percentage of information is also unknown (i.e. it is unknown if the set of known infected nodes represents the entire infected population, or only represents 50% of the full infected set).

The main issue with the partial information case is that *S* cannot be solved directly. The link costs are functions of the transmission probability and timestamps, so Edmond's algorithm cannot be implemented when information (timestamps) is missing, because the necessary link costs cannot be defined *a priori*. This is only a problem when nodes (without information) are included in the tree. If only nodes with information are included in the tree, the original algorithm can be implemented as long as a feasible set of links can be found.

To find the MPST for the partial information case (intended to include potentially infected nodes without known timestamps) a heuristic is proposed, and defined as follows:

i. *Find initial feasible tree rooted at the source node, connecting all the known infected nodes, while including some nodes missing information.
ii. Fill in missing time stamps for all the nodes included in the initial tree (IP).
iii. Find *S* on a sub-network spanning all the nodes in the initial feasible tree using their associated timestamps set in step 2 (IP). (Now with a full set of timestamps link costs can be defined for links connecting all of the nodes included in the initial tree).
iv. Iterate: To further improve the solution (to increase the probability of the tree) iterate between the step 2 and step 3 until convergence is reached.

Using the set of nodes found in step (1), and the optimal set of time stamps set in step (2) which can both be done using an integer program, Edmond's algorithm is implemented on this sub-network in step (3), resulting in *S*. For further improvement, as stated in step (4) the latest MPST can be used as input for step (2), and the initial set of missing timestamps is resolved for. This iteration continues until convergence is reached, at which point the timestamps found in step (2) and set of links found in step (3) no longer fluctuate. This heuristic can be evaluated similarly to the full information case, by calculating $\boldsymbol{Q}$. Currently, the bottleneck in this research is step (1), efficiently identifying an initial tree with guaranteed feasibility. This problem increases in difficulty as less information is made available, and will be an additional topic of future research.

The network structure analyzed in this chapter is representative of human mobility patterns at the community level. This network structure can be derived from transportation systems, such as activity based models. The remaining problems in this dissertation address transportation systems explicitly, representing human mobility directly via air travel networks. The next chapter extends the methodology introduced in this chapter to a macroscopic application implemented on an air travel network. The new objective is to identify high risk travel links in a regional network, responsible for introducing infectious individuals into a previously susceptible region.

CHAPTER 4:

INFERRING INFECTION SPREADING LINKS IN A TRANSPORTATION NETWORK

History has exemplified the significant role of modern transportation in furthering the spread of diseases across cities, states, countries and continents. As previously discussed, recent global epidemics (SARS (2003), Avian Flu (2004, 2005, 2006, and 2007) and Swine flu (2009)) have motivated methodological advancements for integrating inter-regional transportation patterns into previously regional-level disease modeling tools. Today infected humans have the potential to carry viruses into new geographical areas through air travel (as well as other modes such as rail, passenger car, sea. etc.). Additionally, a substantial rise in air traffic has increased the risk of accelerated virus dispersal across geographic distances. This burgeoning risk serves as the main motivation for this research.

The current models for evaluating the impact of air travel on the global spread of contact-based diseases (introduced in the literature review) tend to focus on incorporating travel patterns into large S-I-R agent based simulations. The methodology proposed in this chapter introduces a novel approach for predicting the path of infection (via traveling individuals) between geographic regions. The methodology is implemented on a transportation network where the links capture travel patterns (via air travel) between regions which are represented as nodes.

A method for inferring the most likely path of infection at the local level via a social-contact network was introduced in the chapter 3. Using similar logic, a path of infection can also be inferred between regions. This requires the assumption that an infected individual *y* can only exist in a previously unexposed region, X, if at least one infected individual (either individual *y* or another individual *z*) traveled to X (from a previously infected region) at some previous point in time. Therefore the regional level problem is motivated from the methodology introduced in Chapter 3 developed for the contact network, extended to a new network structure and application.

4.1 PROBLEM DEFINITION

The proposed objective is to identify the links in a transportation network responsible for spreading infection into new, previously unexposed regions. In the contact-network in chapter 3, disease-based transmission probabilities and reported infections are used to infer infectious connections between individuals in a social contact network. The transportation-based problem analogously exploits regional infection data (e.g. day the disease "arrived" at a new location, and daily infection reports) and transportation network properties (set of routes and volume of passenger air travel) to predict the most likely path of infection (*i.e.* set of routes on which infected individuals traveled) that connects all regions which reported infections. In the new network nodes represent regions (cities, states, etc.), and links represent travel routes (flight trajectories, rail connections, etc.). The proposed problem can be viewed as a macroscopic application of the contact-based problem in chapter 3. However the transportation network application poses a new set of challenges.

An example of the proposed model output is shown in Figure 4-1. (This figure represents one *possible* outbreak scenario for a disease that was introduced to the U.S. from Mexico, and proceeded to spread throughout the country). The directed spanning tree (comprised of the set of arrows) branches to the set of reportedly infected states (yellow). In the model output a directed link connecting two regions represents the spread of infection from the "tail" region to the previously uninfected "head" region. One incoming link is chosen for each region to represent the incoming route with the highest probability of carrying an infected traveler/s from a known infected region; thus exposing a new population of individuals to the disease. Using this definition, a link (*i,j*) can only connect *i* to *j* if region *i* was reportedly infected before region *j*. For example the link from Texas to California suggests a traveler from Texas was the most likely source of the disease (later) reported in California. The dark arrow entering Texas identifies Texas as the first infected state in the country. (This map is just an example of the model output, and does not correspond to any actual infection spreading scenario). The most important thing to note about the model output is that is always forms a directed spanning tree, where each region has a single incoming link, but may have multiple outgoing links.

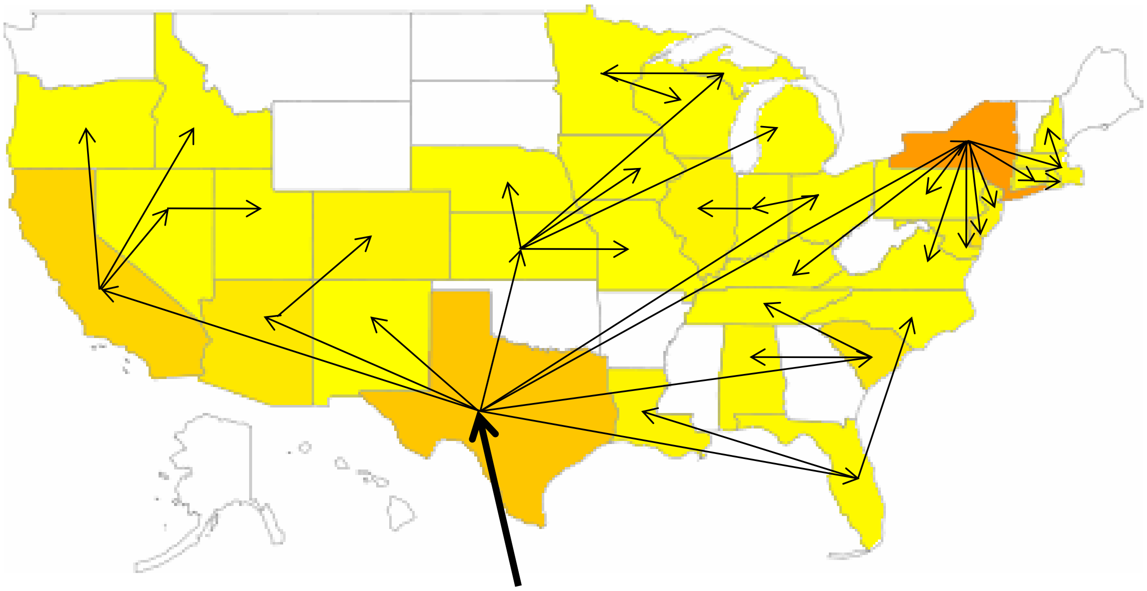


FIGURE 4-1: Example of model output for regional infection spreading scenario

For a given outbreak scenario, the proposed methodology again utilizes network based optimization tools to identify the most likely spreading scenario among a set of regions. This is in contrasts to previously proposed methods (Haydon, 2003) which first enumerate all possible spreading scenarios, and then implement *a posteriori* analysis using various genetic sampling characteristics to identify the most likely scenario from a feasible solution set.

The proposed methodology is intended to identify potentially high risk travel links, and aid in the development of regional level intervention strategies, security measures and surveillance efforts. The results may also provide insight into future outbreak patterns.

4.2 ASSUMPTIONS

To implement the proposed methodology some simplifying assumptions are necessary. Each assumption is listed and expanded upon below.

1. *a priori* knowledge of the underlying transportation network (routes, passenger volume, travel distance, etc.).
2. Temporal infection data is available for the infected regions (e.g. time of initial reported infection).
3. A region can be infected at most once.
4. Infection spreads between regions via infected passengers traveling by air.
5. The outbreak evolved from a single source.

Information required for assumption (1) is available from airlines and government organizations. Issues with assumption (2) may arise when there are multiple reported infections for a single region (which is inevitable for the state-level problem), making it difficult to identify a "timestamp" for the node, which is a necessary input for the model in order to correctly identify a causal relationship between regions. There are multiple options for addressing this issue: 1) Assuming the disease progresses within a population at a constant rate, the peak infection times (this data is available for certain diseases) can be used. Comparing these epidemic peaks would be representative of when the disease was introduced to the region; 2) the time of the first reported infection can also be used, assuming a constant delay in infection reporting across states. Option 2 is chosen for the application included in this chapter.

Assumption (3) demands further clarification as well. Firstly, how do we deal with the case when there is a resurgence of infections in a region which has already been spanned to? It is implicitly assumed that all infections in a region can be traced back to the initial infection in the region, and not a separate source. These later infections could either be a result of a heterogeneous population within the region (the outbreak has traveled through a local contact network to a new community of individuals), or the re-introduction of the disease from a different origin. This possibility of multi-infection at the regional level is analogous to the S-I-S category of diseases at the contact level, where an individual does not require immunity, and can be re-infected. This assumption however introduces the possibility of cycles into the network structure, requiring a different set of modeling tools. For this work the assumption that a region is only infected once is made. Errors associated with this assumption could be minimized by further disaggregating the problem into smaller regions (i.e. from the state to the city).

Assumption (4) is not necessarily unrealistic for larger states such as Texas, or isolated states such as Hawaii or Alaska, however for smaller more dense regions of the country such as in the northeastern U.S. many individuals travel via alternative modes of transportation, and assumption (4) is likely invalid. While this assumption will remain for the application presented in this work, future research should strive to relax this assumption by expanding the network structure to include multimodal human mobility patterns. Human mobility spatial patterns are currently being extensively researched, aided by the availability of cell phone information.

Assumption (5) limits the application of this model to certain outbreaks; however there are scenarios which fall into this category.

4.3 Solution Methodology

Like social networks, the inherent structure of a transportation system makes it an obvious candidate for network modeling tools. The network analyzed in this chapter, $G \in (N, A)$, is defined by a set of nodes, N, which represent regions (i.e. communities, cites, states, countries), and links, A, which connect the regions, representing air traffic patterns. The links $(i,j) \in A$ will have associated weights, $w_{ij}(\cdot)$ that are a function of travel data and infection reports, which will be discussed in detail in section 4.3.4. These weights are intended to represent the relative probability of an infected passenger traveling between regions. Due to the macroscopic level of the transportation network (e.g. disease dynamics among individuals are not accounted for explicitly), this methodology is applicable to a variety of infectious diseases (not just contact based), though still restricted to those originating at a single initial source.

The infection spreading pattern sought forms a directed maximum probability spanning tree. Edmond's maximum branching algorithm (1967) is again implemented, on a sub-network which includes only infected regions, $I \in N$, a feasible link set, $L \in A$, and predefined link costs, P_{ij}, which are a function of the link weights, $w_{ij}(\cdot)$. The set of feasible links $(i,j) \in L$ are those for which region i was reportedly infected before j. In contrast to the contact network problem there is not a restriction on the maximum allowable difference in timestamps at the two ends of a feasible link (this was the infectious period, y, imposed before). While individuals are no longer accounted for explicitly, it is assumed that once infections have been reported in a region, infected individuals continue to reside there, and that region remains a potential threat to those adjacent (connected via air travel) and susceptible (uninfected). Therefore the only constraint on feasible links is $t_i < t_j$, where t_i is the timestamp for node *i.* The resulting maximum probability spanning tree, ***R,*** should include the set of feasible links which branch to every node in *I*, such that the sum of the link costs, $\sum_{\forall(i,j)\epsilon R} P_{ij}$ is maximized. Again, the most computationally intensive portion of Edmond's maximum branching algorithm is the search for and removal of cycles. The assumption that a region can be infected by at most one other region prevents the possibility of a cycle in the outbreak scenario. Therefore the implementation of the algorithm simply requires the following steps:

1. Define the set of feasible links, *L: (i,j)* where $t_i < t_j$
2. Calculate link costs, $P_{ij}(w_{ij}(\cdot))$ for links *(i,j)* in feasible set *L* using the link cost definitions in section 4.3.3.
3. For each infected node, $j \epsilon I$, select the incoming link (*i,j*) with the highest cost, P_{ij}, from the set of feasible adjacent links, $A|j|$

This results in the maximum probability directed spanning tree, *R*. The problem can be formulated as shown below:

$$max \sum_{\forall(i,j)\epsilon\, R} P_{ij}\, x_{ij} \qquad (4\text{-}1)$$

s.t.

$$P_{ij}=f(w_{ij}(\cdot)) \qquad \forall(i,j)\epsilon\, R \qquad (4\text{-}2)$$

$$t_i < t_j \qquad \forall(i,j)\epsilon\, R \qquad (4\text{-}3)$$

$$0\leq w_{ij}\leq 1 \qquad \forall i \in I,\ \forall j \in I \qquad (4\text{-}4)$$

$$\sum_{\forall(i,j)\epsilon\, R} x_{ij} = |I| - 1 \qquad (4\text{-}5)$$

$$\sum_{\forall i\, \in I} x_{ij} = 1 \qquad \forall j \in I \qquad (4\text{-}6)$$

$$x_{ij} = \{0,1\} \qquad \forall (i.j) \in R \qquad (4\text{-}7)$$

$$x_{ij} = \begin{Bmatrix} 1\ if\ edge\ (i.j)\ is\ in\ R \\ 0\ therwise \end{Bmatrix}$$

The objective enforces that the set of links chosen for the spanning tree maximizes the total probability of the tree. Constraints (4-2) to (4-4) pertain to the properties and dynamics of the infection process, while constraints (4-5) to (4-7) enforce the spanning tree structure. Constraint (4-2) defines the link costs. The way the link costs, $P_{ij}= f(w_{ij}(\cdot))$ are defined is one major differentiation between the transportation-network and the contact-network problems. The details of the costs will be discussed in detail in section 4.2.4. Constraint (4-3) defines the set of feasible links, specifically a region *i* can only infect region *j* if *i* is infected first. Constraint (4-4) restricts the link weight, w_{ij} to be fractional. Constraints (4-5) to (4-7) together enforce that the final output is a tree by (4-5) requiring a total of $|I|-1$ links in $\boldsymbol{R}$, where $|I|$ is the number of infected nodes, (4-6) each infected region must have one incoming link, and (4-7) the decision variable, x_{ij}, is binary.

4.3.1 Static vs. Dynamic Model

Two different models are introduced in this chapter: i) static and ii) dynamic. In addition, multiple case studies for each model are defined and evaluated. Results from the two models will be compared across cases. The main difference between the static and dynamic model reduces to their use of infection data. The static model uses *the final outbreak size in each region* while the dynamic model uses *regional daily infection reports*. Infection data is one of many input variables included in the link weight, $w_{ij}(\cdot)$, defined to represent the probability of an infected traveler entering *j* from *i*. The dynamic model therefore defines a time-specific link weight, $w_{ij}^{t}(\cdot)$, to represent the relative probability of an infected traveler entering *j* from *i* at time *t.* The static and dynamic model outputs are now differentiated by R_S or $R_{D,}$ respectively.

For the static model a single iteration of Edmond's maximum branching algorithm is implemented as described above, with the link costs $P_{ij}(w_{ij}(\cdot))$ and feasible link set *L* identified *a priori,* again identifying the incoming link with highest costs $P_{ij}(w_{ij}(\cdot))$ for each infected node. The static model formulation is equivalent to that shown above, where the only difference is the replacement of *R* with $R_{S\,.}$The main issue with the static model is the implicit assumption that the probability of an outgoing traveler being infected (and hereby spreading infection into a new region) is a function of the final number of infections at the route origin, and not the number of infections at the time the traveler departed. This assumption would be valid if the regional level progression of the outbreak was proportional to the final size of the outbreak; however the objective is to predict spreading behavior at the initial stages of an outbreak, at which point this assumption is likely invalid. (As an alternative, the average number of infections over the entire outbreak could also be used to minimize the susceptibility to overestimating infection risk.)

This issue is addressed in the dynamic model, which progressively builds the infection tree, R_D at single time step increments, using real-time infection data. Therefore $w_{ij}^{t}(\cdot)$ is defined, which uses the number of reported infections in a region at time *t* as one input variable in the function. In the dynamic model Edmond's maximum branching algorithm is implemented each time step to identify the incoming route with the highest probability of carrying an infected passenger into a newly infected region, using the dynamic link weights, $w_{ij}^{t}(\cdot)$. The first iteration corresponds to the date the first node in *I* is infected, and the last iteration corresponds to the date the final node in *I* is infected. This time span is represented as *T*. As with the static case the set of feasible links, *L* and time dependent link costs, $P_{ij}^{t}(w_{ij}^{t}(\cdot))$ can be calculated *a priori* (the set of feasible links is equivalent to the set used in the static model). The formal steps are listed below:

1. Define the set of feasible links, *L* (same set as in static model).
2. Calculate time-dependent link costs, $P_{ij}^{t}(w_{ij}^{t}(\cdot))$ for links *(i,j)* in feasible set *L*. For each *t*, this only includes the set of incoming links (*i.j*) for any node *j* s.t. $t_j = t$. The link costs will be defined in section 4.3.3.
3. Starting at *t*=1 (the time the second node is infected), for each infected node, $j \epsilon I$ with timestamps t_j=*t,* identify the incoming link (*i,j*) with the highest cost, $P_{ij}^{t}(w_{ij}^{t}(\cdot))$ from the set of feasible adjacent links, $A|j|$.
4. Repeat Step 3 for *t=T* total iterations, where *T* is the time the last node is infected.

This results in the maximum probability directed spanning tree, R_D. Although the dynamic model requires more link costs calculations *a priori*, this is still a relatively restricted set: For each node *j* the only link costs P_{ij}^{t} required are for the time period node *j* was infected, t=t_j and node pairs including only the set of adjacent nodes $i \epsilon A|j|$ that were infected before *j*, t_i<t_j. Once all the link costs are calculated *a priori*, the maximum branching algorithm is implemented for *T* total iterations, and in each iteration *t* the (feasible) incoming link (*i,j*) with the highest cost, P_{ij}^{t}, is selected for each node *j* (included in the set of infected nodes) with timestamps t_j=*t,*. The dynamic model is therefore able to make predictions based on the real-time status of the outbreak. The only differences in the formal problem definition are i) the use of time dependent link costs

P_{ij}^{t}, 2) a time dependent decision variable, x_{ij}^{t}, and 3) the replacement R_S with R_D. The problem formulation for the dynamic model is shown below:

$$max \sum_{\forall (\mathrm{i,j}) \in R_D} P_{ij}^{t}\, x_{ij}^{t} \qquad (4\text{-}8)$$

s.t.

$$P_{ij}^{t}{=}f(w_{ij}^{t}(\cdot)) \qquad \forall(i,j) \in R_D,\ \ \forall t \in T \qquad (4\text{-}9)$$

$$t_i < t_j \qquad \forall(i,j) \in R_D \qquad (4\text{-}10)$$

$$0 \leq w_{ij}^{t} 1 \qquad \forall i \in I,\ \forall j \in I \qquad (4\text{-}11)$$

$$\sum_{\forall (\mathrm{i,j}) \in R_D} x_{ij}^{t} = |I| - 1 \qquad (4\text{-}12)$$

$$\sum_{\forall i \in I} x_{ij}^{t} = 1 \qquad \forall j \in I \qquad (4\text{-}13)$$

$$x_{ij}^{t} = \{0,1\} \qquad \forall(i.j) \in R_D \qquad (4\text{-}14)$$

$$x_{ij}^{t} = \begin{Bmatrix} 1\ if\ edge\ (i.j)\ is\ selected\ to\ be\ in\ R_D\ at\ time\ t \\ 0\ therwise \end{Bmatrix}$$

Again Constraints (4-9) to (4-11) pertain to the properties and dynamics of the infection process, while constraints (4-12) to (4-14) enforce the spanning tree structure. Although the new decision variable is time dependent, no additional time-based constraints are required (i.e. $\sum_{\forall t \in T} x_{ij}^{t} \leq 1\ \forall\ (i.j) \in R_D$). This is because the algorithm used in the dynamic solution methodology only evaluates each infected node $j \epsilon I$ once, at the time the node is first infected, $t{=}t_j$; and because a node can be infected at most once, there is no way $\sum_{\forall t \in T} x_{ij}^{t} > 1$ for any (i,j).

The main difference between the static and dynamic model is demonstrated using the following example: Assume 1000 total infection cases were reported to have occurred in Texas (throughout the course of the epidemic). The static model then uses this value in $P_{ij}(w_{ij}(\cdot))$ to predict the probability that an infected traveler left Texas for, say, Ohio; which reported its first case one week into the epidemic. However at that time, the actual number of reported cases in Texas was only 15, therefore the link cost used in the static model (with the final infection count of 1000) will overestimate the probability that an infected passenger arriving in Ohio (at that time) came from Texas. The dynamic model instead uses the number of reported infections in each region on a daily basis (in $w_{ij}^{t}(\cdot)$). Therefore the dynamic model identifies the most likely origin of an infected passenger who arrived in Ohio on day 7 of the epidemic, based on the number of reported infections in each previously infected region on day 6 (or within some time window). Ignoring the progressive status of the epidemic may severely limit the predictive capability of the static model.

The static model is however a beneficial tool for prediction, useful for sensitivity analysis, and a good basis for comparison with the dynamic model. And while the dynamic model provides a more realistic prediction, the detailed data required for the dynamic model is not always available and reliable, in which case the best static model should be implemented.

4.3.2 Model Input Variables

The link weights are defined as a function of multiple variables including travel volume, initial infection dates, regional infection counts, travel distance, and regional population. Due to the availability of data, the application chosen for analysis is the 2009 Swine Flu outbreak. Additionally the set of nodes is constrained to include only the

United States and Mexico. The data sources and associated details for each variable are listed below.

i. *Travel Patterns: v_{ij}* is the (average) daily volume of travel between regions *(i,j).* U.S. air traffic data was provided by the Research and Innovative Technology Administration (RITA), a branch of the U.S. Department of Transportation (US DOT), which tracks all domestic and international flights originating or ending in the U.S. and its surrounding provinces (RITA, 2010).

ii. *Infections Timestamps: t_i* is the day of the first official confirmed case in region *i* according to the CDC records. The total time period modeled spans the day of the first reported infection (by State) which was April 21 in California, to the day the last state confirmed (U.S.) case which was the U.S. Virgin Islands on June 16th. This is a total of 57 days (e.g. *T*=57). The same set of timestamps applied to the static and dynamic model, and therefore the set of feasible (potential infection spreading) links also remains constant between models.

iii. *Regional Infection Counts*: The infection data set is provided by the Center for Disease Control and Prevention (CDC) 2009 H1N1 website (http://www.cdc.gov/h1n1flu/), and quantifies the daily progression of the outbreak.

  a. *Dynamic Model Variables*: The dynamic model uses the State-level number of reported infections per day, o_{it}, from April 21st to May 14th. Because o_{it} values are required for the entire time period, *T* (to predict the cause of infection for States infected later)*,* after May 14th the number of reported infections in each state is left constant. This approximation is a minor issue because there are only five states still uninfected after May 14th: Puerto Rico, Alaska, Wyoming, West Virginia, and the U.S. Virgin islands.

  b. *Static Model Variables***:** The same data set is used for the static model, but only the final infection count (as of May 14th) is required, which is set equal to the variable $o_{i.}$

iv. *Regional Population*: p_i is the population of region *i*, provided by the 2010 U.S. Census Bureau.

v. *Travel Distance:* d_{ij} is the average travel distance between regions, calculated in ArcGIS, an integrated Geographic Information Systems (GIS) software package. The average distances are computed for each route as the geodesic distance between the geographic centers of each region, using latitudinal and longitudinal coordinates.

To help familiarize with the data sets listed above, Table 4-1 lists the top ten ranked states for five different variables used in the link weight functions:

i. Timestamp (order of infection), t_i
ii. Official reported infection counts, o_i
iii. Total outgoing travel volume for each state i, $\sum_j v_{ij}$
iv. Population, p_i
v. Ratio of reported infections to population, o_i/p_i

Familiarization with the data aids in the interpretation of the results. For example, it is immediately obvious that some states rank very high in multiple categories, (e.g. Mexico, California, Texas, New York, Illinois and Pennsylvania), therefore the model will likely predict that these states play a major role in the spread of disease. The states with the highest ranking for o_i/p_i vary the most from the other rankings shown. This ranking should be accounted for in the analysis, (i.e. a highlighted role of Utah, New Mexico, Oregon or Iowa in case IV could be explained by their relatively high o_i/p_i ranking).

TABLE 4-1: State Rank in terms of Variables

	State Ranking				
Rank	**Timestamp**	**Infection Count**	**Outbound Travel**	**Population**	**Infection/pop**
1	Mexico	Mexico	California	Mexico	Wisconsin
2	California	Illinois	Texas	California	Delaware
3	Texas	Wisconsin	Florida	Texas	Arizona
4	Kansas	California	Georgia	NewYork	Illinois
5	NewYork	Texas	Illinois	Florida	Washington
6	Ohio	Arizona	New York	Illinois	Utah
7	Indiana	NewYork	Colorado	Pennsylvania	New Mexico
8	Arizona	Washington	North Carolina	Ohio	Oregon
9	Maine	Michigan	Nevada	Michigan	Iowa
10	Massachusetts	Massachusetts	Arizona	Georgia	Mexico

4.3.3 Link Weights

For both the static and dynamic models, a link weight representative of the regional "transmission" probability $w_{ij}(\cdot)$ is required. In the social-contact network in chapter 3 the link weight is equivalent to the transmission probability. In the regional-level problem such disease properties should be implicitly considered, but cannot be applied directly. The regional macroscopic level problem requires a link weight that is representative of the interaction between two regions (*i.e.* passenger travel volume), while also accounting for the probability of those passengers being infected (and therefore capable of spreading the disease). Therefore w_{ij} should be a function of the attributes of the two regions (e.g. population size), the size of the outbreak in a given region (e.g. number of reported infections), human mobility patterns (e.g. air traffic volumes), and perhaps other link specific characteristics such as travel distance. This function must be defined *a priori* because it is necessary input for the spanning tree algorithm, and directly determines R_S and R_D. As the historical data necessary to calibrate this type of model does not exist, the goal is instead to explore a variety of link costs which vary based on their functional form and input variables, and evaluate the different model outputs. The goal is to find a link weight function which correctly predicts the regional infection spreading pattern.

The remainder of this section introduces the various proposed link weights, $w_{ij}(\cdot)$ and $w_{ij}^{t}(\cdot)$ by case number. The only variation from the static to dynamic weight function is the replacement of o_i with o_{it}, the dynamic infection data. For now the same approximated daily travel volume, v_{ij} is used for both the static and dynamic model, and the population for a region, p_i is assumed to remain constant over the course of the outbreak. Therefore each case introduced is applicable to the static and dynamic model; however the dynamic model will often results in different predictions due to the variation in infection data provided. In each of the link weight functions the regional population size is factored by 10,000. This factor is used because the ratio of infections to population is extremely low. The factor is selected such that $w_{ij}(\cdot)$ and $w_{ij}^{t}(\cdot)$ always falls within the range (0,1). The following notation sums up the list of variables used in the remainder of this chapter:

TABLE 4-2: List of Variables

$w_{ij}(\cdot)$	Weight assigned to link (*i.j*) for the static model
$w_{ij}^{t}(\cdot)$	Weight assigned to link (*i.j*) for the dynamic model
P_{ij}	Cost assigned to link (*i.j*) for the static model
P_{ij}^{t}	Cost assigned to link (*i.j*) for the dynamic model
T	Total time span of outbreak
t	Time period (day) during outbreak
t_i	Timestamp for node *i*, representative of the date of the first confirmed case in region *i*
v_{ij}	Number of passengers traveling on route (*i,j*) at time *t*
o_i	Number of reported infections in region *i*, used in the static model
o_{it}	Number of reported infections in region *i* at time *t*, used in the dynamic model
p_i	Population of region *i*. This can be assumed to remain constant.
d_{ij}	Travel distance between regions *i* and *j*
q_t	Probability of a random traveler being infected, set equal to o_{it}/p_i
n_t	Number of passengers traveling on route (*i,j*) at time *t*; v_{ij} is used as an approximation

4.3.3.1 Case Studies

Case I, defined as $w_{ij}(I)$, represents the simplest link weight function considered, and includes only one variable, *passenger travel volume*, which intuitively plays a significant role in the probability of spreading infection between regions. In this case the link weight is simply proportional to the normalized travel volume on a route: the route travel volume divided by the maximum travel volume over all routes. This function is used as a base case, and clearly does not account for attributes specific to the outbreak or region, and is therefore likely an unrealistic estimate for spreading infection. It is however useful for assessing the role of additional variables. These additional variables will be incorporated into the other cases sequentially.

Case II introduces *regional population* into the link weight function. The role of population is revealed by directly comparing $R_{S,D}(I)$ and $R_{S,D}(II)$. In addition both case I and II will not vary between the static and dynamic models because the link weights are not a function of the infection data, which is the only input variable that differs between the two models. In case II population is included in the denominator with the intention of minimizing model bias towards largely populated regions. From table 4-1 it is apparent that states with large outbound travel volumes are also among the most populated states. Two versions of Case II were evaluated: *II.i*) population is included in the numerator, $w_{ij}(II.i) = v_{ij} * P_i$ and *II.ii*) denominator, $w_{ij}(II.ii) = \frac{v_{ij}}{P_i}$. As expected case *II.i* provides highly similar predictions to case *I* because of the positive correlation between travel volume and state population size. Only results from case *II.ii* are presented in this chapter.

Case III is the first to introduce properties of the outbreak into the link weight function by including the number of reported infections at the route origin. Case III uses the product of the normalized travel volume on a route and the number of reported infections at the route source. Comparing this case directly with case I can provide insight into the role of infection data on the model predictions.

Case IV is representative of the probability an infected individual will travel between origin i and destination j by multiplying the proportion of the population that is infected, $o_i/p_i,$ with the passenger travel volume v_{ij}. This is not the exact probability of an infected individual leaving i, or traveling to $j,$ due to the approximated input variables (infection count and estimated daily travel volume), but is representative of the event, relative to other routes. In addition an implicit assumption made in this research is first introduced here; *an infected and healthy passenger is equally likely to travel*.

Case V incorporates all the same variables as case IV; however the function is derived using the binomial probability distribution, where $w_{ij}(V)$ defines the probability of at least one infected passenger traveling between regions. The function has a more intuitive explanation using the dynamic model. Again assuming that infected and healthy individuals are equally likely to travel on a given day, the probability of an infected traveler leaving a region on a particular route (on a particular day) can be defined using a binomial probability, $b(k_t;n_t,q_t)$, where k_t is the expected number of infected travelers on the route on a given day, n_t is the total number of daily travelers on a given route, and q_t is the probability of a random traveler being infected on that day. For simplification, the t subscript will be left out of the formulation in the following explanation. The probability of at least one infected traveler traveling on a given day is the complement of no infected travelers, $b(k>0;n,q) = 1-b(0;n,q)$. By recalculating these probabilities each day (or each time step in the algorithm), the most likely source of infection for a newly infected region can be chosen based on the current status of the outbreak (*i.e.* which regions are already infected).

The probability of no infected travelers being on a particular route (i,j) reduces to:

$$b(k=0;n,q) = C(n,k)q^k\ (1-q)^{n-k} = (1-q)^n \tag{4-15}$$

Therefore the probability of at least one infected passenger is:

$$b(k>0;n,q)=1-\ b(k=0;n,q)=1-(1-q)^n \tag{4-16}$$

Plugging in the problem variables:

$$b(k>0;n,q)=\ 1-(1-q)^n\ =1-(1-\frac{O_{it}}{P_i})^{v_{ij}} \tag{4-17}$$

Equation (4-17) is the resulting link weight function used in the dynamic model, $w_{ij}^{t}(V)$. The same logic and formulation are used in the static model, but the updated infection data is ignored, and o_{it} is replaced with o_i. While case V has the potential to be the most accurate prediction measure; with the approximated input data used in this problem the predictions are not guaranteed to reflect the actual risks.

The next three cases incorporated travel distance into the link weights. Case VI and VII build up to case VIII, which is inspired by the gravity model commonly used in physics and transportation applications. Case VI simply divides travel volume over travel distance; resulting in a higher cost (which translates to a higher probability of infection) for routes with higher travel volumes, and shorter travel distances. Case VI does not include infection data, so the static and dynamic results will not vary.

Case VII is an extension of case VI, introducing infection data, and is simply the product of $w_{ij}(V)$ and the infection count, o_i. Case VII and case VI can be directly compared to reveal the role of infection data in R.

Case VIII is inspired by the general gravity model used in transportation theory which assumes the commuting flow, f_{ij} between regions i and j is proportional to the population in each region, P_i and P_j, the distance between the two regions, d_{ij}, and some proportionality constant, C. The gravity model is representative of the attraction between all pairs of regions. A similar logic is used in this work to represent the "attraction" to a given destination experienced by an infected passenger. In case VIII the weight for route (i,j) is a function of the number of reported infections at i, the travel volume v_{ij}, the population of each region, divided by the travel distance. All variables except for travel volume are multiplied in the numerator, resulting in increased link costs for the more highly traveled routes, populated regions, and reported infections. The only variable in the denominator is the travel distance, resulting in a higher link cost for shorter travel distances, all other things being equal. Because the algorithm chooses a predecessor region for each infected destination, including the destination population, p_j in the numerator does not have any impact because it increases all the incoming link costs proportionally. In summary, the link weight functions for each case are shown in Table 4-3.

TABLE 4-3: Link Weight Functions

Case	Static	Dynamic
I	$w_{ij}(I) = \frac{v_{ij}}{\max_{ij} v_{ij}}$	$w_{ij}^{t}\ (I) = \frac{v_{ij}}{\max_{ij} v_{ij}}$
II	$w_{ij}(II) = \frac{v_{ij}}{P_i}$	$w_{ij}^{t}(II) = \frac{v_{ij}}{P_i}$
III	$w_{ij}(III) = \frac{v_{ij}}{\max_{ij} v_{ij}} * o_i$	$w_{ij}^{t}(III) = \frac{v_{ij}}{\max_{ij} v_{ij}} * o_{it}$
IV	$w_{ij}\ (IV) = \frac{o_i}{P_i} * v_{ij}$	$w_{ij}^{t}\ (IV) = \frac{o_{it}}{P_i} * v_{ij}$
V	$w_{ij}\ (V) = 1\text{-}(1 - \frac{o_i}{P_i})^{v_{ij}}$	$w_{ij}^{t}(V) = 1\text{-}(1 - \frac{o_{it}}{P_i})^{v_{ij}}$
VI	$w_{ij}\ (VI) = \frac{v_{ij}}{D_{ij}}$	$w_{ij}^{t}(VI) = \frac{v_{ij}}{D_{ij}}$
VII	$w_{ij}\ (VII) = \frac{v_{ij}}{D_{ij}} * o_i$	$w_{ij}^{t}(VII) = \frac{v_{ij}}{D_{ij}} * o_{it}$
VIII	$w_{ij}\ (VIII) = \frac{o_i * P_i * P_j * v_{ij}}{D_{ij}}$	$w_{ij}^{t}(VIII) = \frac{o_{it} * P_i * P_j * v_{ij}}{D_{ij}}$

4.3.4 Link Costs

Now that $w_{ij}(\cdot)$ and $w_{ij}^{t}(\cdot)$ are defined, the same logic as in the contact network can be applied to generate link costs, P_{ij}, for input to the algorithm. The formulations are provided in table 4-4. For the social-contact model the link cost function, P_{ij}, accounts for the infection delays, or the opportunities node *i* had to infected node *j* but did not, and the probability infection occurred between two individuals once. Analogously, $(1- w_{ij}(\cdot))$ is the probability that infection did not occur between two regions once region *i* was infected, and $w_{ij}(\cdot)$ is the probability that infection occurred once. Specifically in this problem the probability of delay, $(1- w_{ij}(\cdot))$, represents the case where an infected passenger had the opportunity to travel to a new region and spread infection, but did not. These opportunities are defined as the days between the first reported infection in region *i* and the first reported infection in region *j*, during which daily travel routes were ongoing between *i* and *j*. The same logic applies to the dynamic model; however the link weights are recalculated each time period *t*.

A possible outcome from using link cost function *P* is that if a high traffic route does not result in infection for a specific region on the first opportunity made available, it is unlikely to be chosen as the cause in infection at a later time. This is because when $w_{ij}(\cdot)$ is nearly one, the probability of the infection occurring on the second or third chance is extremely low, $(1- w_{ij}(\cdot)) \approx 0$. This effect becomes evident when the time lapse between reported infections in two regions is overestimated (perhaps because of faulty data, or late reporting), in which case the state that is the likely source of infection is not likely to be correctly identified by the model.

To mitigate this issue, the methodology is also implemented for link costs, *w*: $P_{ij} = w_{ij}(\cdot)$ and $P_{ij}^{t} = w_{ij}^{t}(\cdot)$ directly, which ignores the impact of infection delay completely. The set of feasible links still remains the same. Using *w*, the models will likely identify more causal routes where the calculated link weights are maximized, even if the initial reported infections (timestamps) are further apart. Neglecting the time gap between initial infections presents its own issue, but is a useful tool for comparison. The link costs, P_{ij} and P_{ij}^{t} are defined for the static and dynamic model as in Table 4-4:

TABLE 4-4: Link Cost Functions

	Static	Dynamic
P	$P_{ij} = (1- w_{ij}(\cdot))^{(\Delta t-1)} *(w_{ij}(\cdot))$	$P_{ij}^{t}= (1- w_{ij}^{t}(\cdot))^{(\Delta t-1)} *(w_{ij}^{t}(\cdot))$
W	$P_{ij} = w_{ij}(\cdot)$	$P_{ij}^{t}= w_{ij}^{t}(\cdot)$

For simplification, the output specific to each model (static or dynamic), case number (I-V) and link cost (P or w) combination will be represented as $R_{model}^{link\ cost}(case)$ or $R_{M}^{L}(C)$. For example $R_{D}^{w}(V)$ represents the output tree for the dynamic model, case V, when the link cost $P_{ij}^{t}= w_{ij}^{t}(\cdot)$, while $R_{D}^{P}(V)$ represents the output tree for the dynamic model, case V, when the link cost $P_{ij}^{t}= (1- w_{ij}^{t}(\cdot))^{(tj-ti-1)} *(w_{ij}^{t}(\cdot))$.

4.4 Network Structure

The network analyzed in this chapter is limited to the United States and Mexico, where the majority of infections were concentrated during the initial stages of the 2009 Swine flu. The network includes 53 nodes representing the 50 United States, the U.S. Virgin Islands, Puerto Rico, and Mexico, and a set of directed links connecting all the regions with direct air travel. Only links from Mexico into the U.S. and links originating and ending in the U.S. are included in the network. No traffic exiting the U.S. is accounted for at this point. The network is created using the air traffic data provided by RITA (2010). Using monthly passenger (airport-to-airport) travel volumes, the state level travel volumes were calculated by aggregating passenger market data across all airports in a given state. The same aggregation was used to consolidate all travel out of Mexico into any State. The resulting data set includes the total monthly passenger travel volume from Mexico into each State, and all domestic state-to-state travel in May 2009. May was chosen because it was closest to the peak of the outbreak. The daily travel volume is approximated by factoring the total monthly travel volumes by 31 (approximated travel days in May). This data set is intended to represent single scale (e.g. air travel) human mobility between regions in the time period of the outbreak. The reason for aggregating travel volumes to the state level is based on the state-level availability of a complete infection data set. If city level infection data was available for all cities in the country, then the same methodology could be applied to the disaggregated problem, tracking infection between cities. The city level model would likely be a better platform for tracking the infection across space and time. This introduces an additional motivation for this work; highlighting the need for improved infection data collection efforts, and the potential benefits for making it available to researchers. Lastly, the transportation data used in this paper focus on passenger travel volumes and does not include cargo flights.

An example of an air transportation network is shown in Figure 4-2 illustrating Continentals daily service patterns within the U.S. (NOTE: Figure 4-2 is solely Continental's travel patterns, and not the actual network structure created after aggregating travel across all airlines, or aggregating up to the state level.) The most obvious characteristic of this network is the hub and spoke structure, which is consistent with a power law degree distribution. The network structure plays an integral role in the dissemination of a disease throughout the country. Specifics concerning the role of the network structure were highlighted in Chapter 3, within the context of a social contact network.

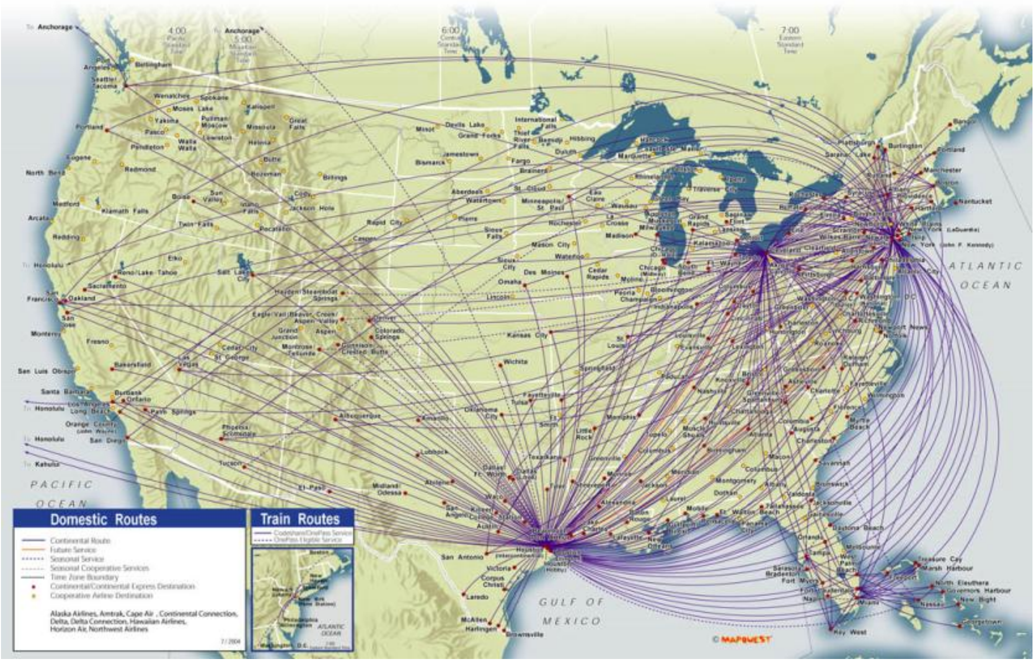

To Anchorage
4:00 Pacific Standard Time
5:00 Mountain Standard Time
6:00 Central Standard Time
7:00 Eastern Standard Time
PACIFIC OCEAN
ATLANTIC OCEAN
GULF OF MEXICO
To Honolulu
To Kahului
Domestic Routes
Continental Route
Future Service
Seasonal Service
Seasonal Cooperative Services
Time Zone Boundary
Continental/Continental Express Destination
Cooperative Airline Destination
Alaska Airlines, Amtrak, Cape Air , Continental Connection, Delta, Delta Connection, Hawaiian Airlines, Horizon Air, Northwest Airlines
7/2004
Train Routes
Codeshare/OnePass Service
OnePass Eligible Service
Boston
Philadelphia
Wilmington
ATLANTIC OCEAN
Washington, D.C.
© MAPQUEST

FIGURE 4-2: Example of U.S. Air Traffic Network (Continental, 2010)

The transportation network used for the Swine Flu application aggregates all domestic and international carriers operating within the U.S. to the state level. The final network has 53 nodes (the 50 U.S. states, U.S. Virgin Islands, Puerto Rico and Mexico), 1829 links and carries over 53 million passengers. Before aggregating all routes to the state level, the (airport level) air traffic network had 17,484 links. The most significant impact of aggregating the network to the State-level is the resulting change in network structure. If only the existence of links is accounted for (no link weights for passenger volume), the hub and spoke degree distribution of the airport-airport network is simplified to a more uniform state-to-state network structure. For this network structure a link exists between two regions if there is any amount of travel between those regions (travel volumes are not explicitly accounted for). The uniformity of the State-level degree distribution is illustrated in the Figure 4-3 below. The number of nodes with a given degree increases nearly linearly with the degree. As illustrated in chapter 3, the network structure plays a significant role in the performance of the solution methodology proposed; the resulting network properties must be accounted for when dealing with the aggregated network.

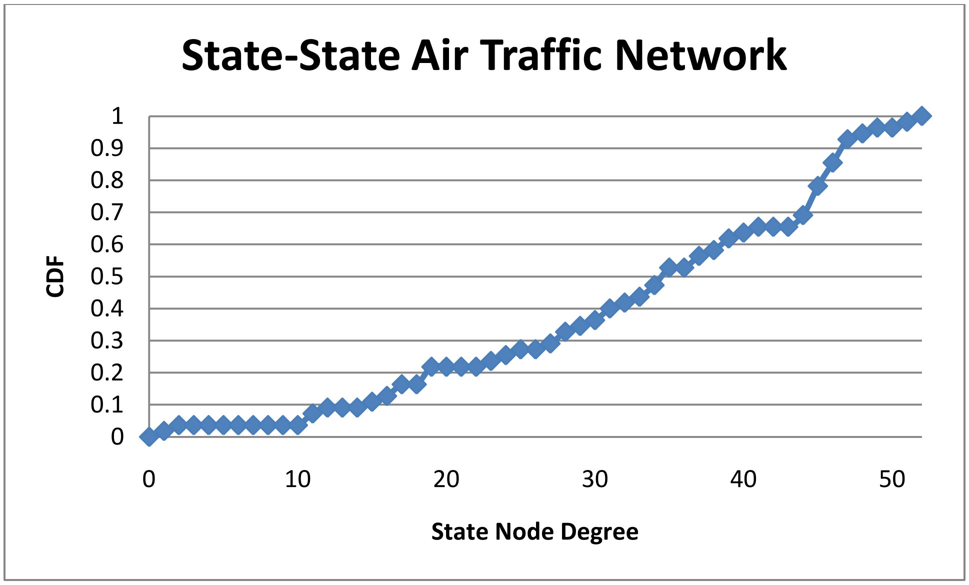


FIGURE 4-3: Degree distribution for State level aggregated air traffic network

To further explore properties of the aggregated air traffic network the links are weighted by their associated passenger volume, rather than just defined by their existence. This weighted network structure takes on a different form. To illustrate this each link (*i,j*) is assigned a value, q_{ij}, equal to the travel volume on that link divided by the total travel volume across all routes, $q_{ij} = \frac{v_{ij}}{\sum_{ij} v_{ij}}$,. The passenger-volume weighted network structure (Figure 4-4) more closely resembles a power law network. The state indices are ordered in increasing passenger volume. For this weighted network eight states handle 50% of the total passenger volume for the country, with California alone handling 10% of the total passenger volume, and most states handle far less than 10%.

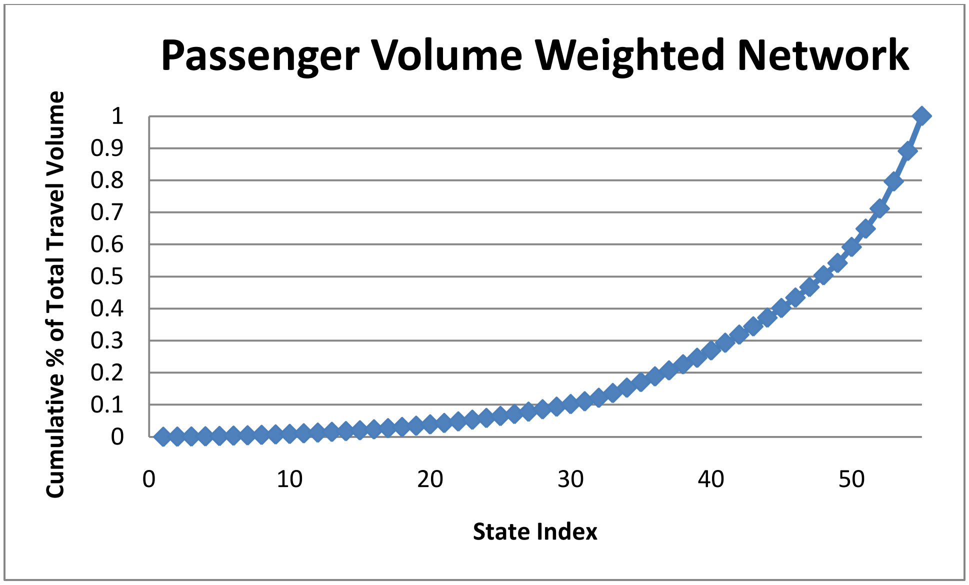


FIGURE 4-4: Passenger-volume weighted network structure

The network structure analyzed for the Swine Flu application is aggregated to the state level. This level of aggregation results in a network structure that is not the typical power law network commonly associated with air traffic networks, but instead a more uniform structure. For this network structure the link weights are defined as a function of travel volume, among other factors which result in a heterogeneous set of link weights.

The combination of these properties, a uniform network structure and heterogeneous link costs, factors into the model performance.

4.5 Measure of Performance

In addition to defining the link costs, another major challenge is evaluating the model's performance. The "best" model should ideally predict the actual spreading pattern that occurred, which is unknown. One evaluation measure is to compare ***R*** against traveler patient survey data collected to identify the most likely source of the disease. However a complete data set of infection-causal travel routes is not currently available. The lack of available information makes it difficult to assess the validity of the proposed model.

An alternative method of evaluation is to compare ***R*** to link-level predictions from other published models developed to predict the same outbreak scenario. One such family of models uses Phylogeographic analysis, a common approach in molecular ecology, connecting historical processes in evolution with spatial distributions [Knowles, 2002]. A model specific to Swine Flu was published by Lemey *et.al.* (2009), and infers the phylodynamic spread in time and space of the virus by employing a recently developed Bayesian statistical inference framework. The process involves modeling spatial diffusion on time-measured genealogies as a continuous-time Markov chain (CTMC) process over discrete sampling locations, resulting in link-based predictions between geographic regions. The model using well established sequence evolution models (Drummond, 2007) along with phylogenetic likelihood evaluation (Suchard, 2009). This procedure leads to a set of regional links that appropriately explain the spatial-temporal process, complimented with a formal Bayes Factor (BF) test of the significance of the linkage between locations. Rates yielding a BF > 3 revealing epidemiological linkages between United States are provided in Table 4-5. The links are listed such that the "from" timestamp is always less than the "to" timestamp, presented in this way for ease of comparison with the proposed model output. In Table 4-5 the strongest link is observed between Mexico and Texas. Mexico is involved in only two additional links with BF > 5, Illinois and Florida. The earliest dispersal event between Mexico and California is not supported because it precedes the time frame of the analysis.

TABLE 4-5: Results from Lemey (2009) for H1N1 Phylogenetic Analyses

Rank	From	To	BF
1	Mexico	Texas	65.92
2	Kentucky	Virginia	38.99
3	Colorado	South Dakota	25.11
4	Indiana	Pennsylvania	19.08
5	Arizona	Nevada	11.51
6	Wisconsin	Louisiana	9.45
7	Ohio	Arizona	9.24
8	Kentucky	Utah	8.91
9	Indiana	Michigan	7.03
10	Illinois	Pennsylvania	6.67
11	Missouri	South Dakota	6.42
12	Utah	Montana	6.36
13	Mexico	Illinois	6.28
14	Kentucky	Montana	5.90
15	Arizona	Tennessee	5.79
16	Virginia	Utah	5.72
17	Mexico	Florida	4.97
18	Indiana	Illinois	4.94
19	Missouri	Illinois	4.57
20	Ohio	Tennessee	4.34
21	Michigan	Pennsylvania	4.28
22	Florida	Missouri	3.83
23	Florida	Tennessee	3.69
24	Arizona	Florida	3.39
25	Virginia	Montana	3.23
26	California	Tennessee	3.20
27	Nevada	Virginia	3.19
28	Florida	Illinois	3.07

Although the links identified in Table 4-5 do not account for factors such as human mobility or a transportation network structure, these results are one of the only published results coupling regions with respect to the H1N1 outbreak. This analysis provides one basis for comparison with the proposed models, although relying solely on these results is insufficient for multiple reasons: i) Lemey's results are probabilistic, so comparable results are not proof of accuracy, especially for the links with low BF values, ii) the pairings in Table 4-5 are not directional, so they are not equivalent to the causal-links representing directed human transport in the proposed model, iii) Only a subset of infected states are included in Lemey's model due to a restricted data set, and insufficient evidence for pairing two regions, and iv) for the states included in the results, there are multiple incoming links for a given node, the results do not form a tree structure. None the less, the model provides the only known link-based predictions which are comparable to the type of predictions by the proposed model. Therefore as one measure of performance $\boldsymbol{R}$ will be compared with inferred epidemiological links from Lemey (2009).

Another option for evaluating the model is to measure the robustness of the proposed methodology, by quantitatively comparing $R_M^L(C)$ for the different model-case-link costs combinations. $\boldsymbol{R}$ is constructed by selecting the incoming travel route (for each state) most likely to carry an infected individual. In a tree structure each node has a single incoming link therefore each $R_M^L(C)$ is presented as a list of directed links (representing interstate travel routes that most likely spread infection). Using table formatting, the "from" state is listed in the left hand column and the "to" state in the right hand column. The links are ordered alphabetically by destination state name. For each $R_M^L(C)$ the second column (set of "to" nodes) remains constant, as this is just the set of infected regions. This presentation format allows a direct comparison of the "from" node list for each M-L-C combination evaluated. Therefore the robustness is calculated as the percentage of predecessor nodes shared among models. This measure of comparison is only applicable for a direct comparison of two $R_M^L(C)$ trees, though any combination of two can be evaluated.

4.6 Numerical Results and Analysis

Using the link-based presentation format described previously, results are provided in tables 4-6 to 4-21 for each M-C-L combination evaluated. An analogous set of results is presented for the link cost function *w*: $P_{ij} = w_{ij}(\cdot)$, (Tables 4-6 to 4-13) which ignores the time delay, and *P*: $P_{ij} = (1 - w_{ij}(\cdot))^{(tj-ti-1)} *(w_{ij}(\cdot))$, (Tables 4-14 to 4-21) for which time delays are accounted for. (The tables are listed at the end of the chapter.) For each link cost function the tables are ordered by case number (I-VIII), with the static results on the left, and dynamic on the right. To illustrate the commonality between $R_S^L(C)$ and $R_D^L(C)$, the set of links predicted by both the static and dynamic model for a given case are highlighted in yellow. This presentation format provides a visual interpretation of the model robustness. In general the "overlap" between different cases, models and link cost functions will be defined as the percent of links shared between two $R_M^L(C)$ trees.

The overlap between the proposed model output and phylogenetic analysis by Lemey (2009) is illustrated as well, with the links highlighted in red. Although Table 4-5 lists 28 links that are feasible for the proposed model, the maximum number of links that may be shared with any $R_M^L(C)$ tree is 14. This is because only one incoming link is selected for each node in $\boldsymbol{R}$, and only 14 "destination" states have predicted predecessors in table 4-5. For example, four (previously infected) regions are connected to Illinois;

from Mexico, Indiana, Missouri and Florida. Because each node can only have one predecessor in the proposed model, at most one of these links can be identified.

Lastly, an inconsistency with the data sets exists, which introduces an issue to be addressed in future research. Delaware has incoming flights from only one state, North Carolina, and North Carolina is reported as becoming infected after Delaware. Therefore the model is unable to predict a predecessor for Delaware. While this issue could be attributed to incorrect data it is also possible that the cause of infection in Delaware was introduced through some alternative mode of transportation. The same possibility applies to many of the northeastern states located within close proximity, with multiple modes available for inter-state travel (i.e. rail, auto). Accounting for alternative modes of travel is future research topic that will be expanded upon in the conclusions. However, at this stage the assumption that infection between states only occurs via air travel remains. Currently Mexico is listed as the default source of infection for Delaware.

4.6.1 Case Specific Results

Cases I and II and IV immediately stand out because there is no difference between the static and dynamic model predictions. This is illustrated in tables 4-6, 4-7, 4-11 and 4-14, 4-15 and 4-19 which are fully highlighted in yellow, implying that the static and dynamic models predicted the exact same outbreak scenario. More formally:$R_S^P(I) = R_D^P(I), R_S^P(II) = R_D^P(II),\ \ R_S^P(VI) = R_D^P(VI),\ \ R_S^w(I) = R_D^w(I),\ \ R_S^w(II) = R_D^w(II)$ and $R_S^w(VI) = R_D^w(VI)$. The 100% overlap between the static and dynamic model for cases I, II and VI is expected because the link weights, timestamps and set of feasible links are the same for the static and dynamic models The link weights are identical because no infection data is accounted for, which is the only input variable that differs between the static and dynamic model.

Another interesting observation is the comparison between cases I and II, which share 50% of their links. This implies that including population size in the link weight denominator does impact the model predictions. In general some of the larger "from" states in case I are replaced with less populated states, highly trafficked states in case II. The results from Case II under both link cost functions share more predictions with Lemey (2009) than Case I.

Case III is the first function to introduce the outbreak dynamics into the cost function. By comparing $R_M^L(I)$ and $R_M^L(III),$ the model sensitivity to infection counts is revealed. When the time delay is ignored there is 62% overlap between $R_S^w(I)$ and $R_S^w(III),$ and 49% overlap between $R_D^w(I)$ and $R_D^w(III),$ suggesting the static model is less sensitive to infection counts. For the link costs, P, both the static and dynamic model are more sensitive to infection data, both $R_S^P(I)$ and $R_S^P(III)$ and $R_D^P(I)$ and $R_D^P(III)$ share 40% and 35% of their links respectively. A direct comparison between $R_S^w(III)$ and $R_D^w(III)$ is illustrated in table 4-8 in which 62% of the links overlap; while only 35% of the links overlap $R_S^P(III)$ and $R_D^P(III)$, illustrated in table 4-16. The variation between the static and dynamic models is expected because the dynamic model uses daily infection reports in updated link weights, which should result in different predictions if the daily infection counts are not exactly proportional to the final infection counts used in the static model. In addition, accounting for time delays exaggerates the differences in the infection data.

Case IV also includes travel volume, infection counts, and population in the link weight function, again including population in the denominator. Although Cases III and IV differ only in the population variable, they do not overlap highly in predictions, with the exception of $R_D^w(III)$ and $R_D^w(IV)$ which share 84% of their links. The dynamic model is expected to be more sensitive to the infection data because of the temporal properties of the problem. As with Case I and III, Cases II and IV can be directly compared to evaluate the role of infection data, and are found to only share 31%-34% of their links, dependent on the specific $R_M^L(C)$ tree. This suggests the role of infection data in the model predictions is more significant when the regional population is included. In terms of comparing the static and dynamic models directly, 45% of the links overlap between $R_S^w(IV)$ and $R_D^w(IV)$, and 35% between $R_S^P(IV)$ and $R_D^P(IV)$. The minimal overlap between the static and dynamic models suggests that the final infection counts used in the static model are not proportional to the daily infection counts.

Case V uses a link weight which approximates the binomial probability that at least one infected passenger is on a given route, using the same set of input variables as case IV. This similarity is exemplified; $R_S^w(IV)$ and $R_S^w(V)$ and $R_D^w(IV)$ and $R_D^w(V)$ share 100% and 71% of their links respectively. However for the link cost function, P, the results are significantly different; $R_S^P(IV)$ and $R_S^P(V)$ and $R_D^P(IV)$ and $R_D^P(V)$ are only in compliance 50% and 13% of the time. These results illustrate the model sensitivity to the functional form of the link weight. In comparing the static versus dynamic model for case V, $R_S^w(V)$ and $R_D^w(V)$ share 36% of their links, while $R_S^P(V)$ and $R_D^P(V)$ share 32%. This furthers the observable trend differentiating the two models when temporal infection data is included. This also suggests a single infection count for a region is not necessarily an adequate measure of the risk that region poses to other susceptible regions though out the outbreak.

Case VI is the first to introduce distance into the link weight. As in case I and II, case VI does not include infection data in the cost function so the static and dynamic models will predict the same set of links. Comparing cases VI and I directly demonstrates the role of distance in the cost function, with travel volume being the only other variable. For link cost w both $R_S^w(V)$ and $R_S^w(I)$ and $R_D^w(V)$ and $R_D^w(I)$ share 74%, while for link cost P both $R_S^P(V)$ and $R_S^P(I)$ and $R_D^P(V)$ and $R_D^P(I)$ share 65%. This is relatively high, especially for link cost w, suggesting distance does impact the predictions, but not as significantly as travel volume. However no other variables are included at this stage, so the predictions are not likely to be accurate.

Case VII extends case VI, and includes infection data in the cost function. These two cases can be compared (as in case I and III) to explore the role of infection data on R. It was found $R_S^w(VI)$ and $R_S^w(VII)$ share 65%, and $R_D^w(VI)$ and $R_D^w(VII)$ share 59%. $R_S^P(VI)$ and $R_S^P(VII)$ share 65%, and $R_D^P(VI)$ and $R_D^P(VII)$ share 44%. The difference is larger on average than the difference between case I and III, but shares similar ranking characteristics. In comparing the static and dynamic model directly for case VII, as we have seen before, the static and dynamic model differ more under link cost P, where $R_S^w(VII)$ and $R_D^w(VII)$ share 72% of their links, and only 54% overlap between $R_S^P(VII)$ and $R_D^P(VII)$.

Case VIII introduces population into the link costs. The static and dynamic models differ only mildly for case VIII, 76% for both $R_S^w(VIII)$ and $R_D^w(VIII)$ and $R_S^w(VIII)$ and $R_D^w(VIII)$. For both cases VII and VIII the only links shared with results from Lemey are outgoing from Mexico. An interesting comparison is between the pseudo-binomial case V, and the gravity inspired case VIII. The predictions are illustrated graphically in figures 4-5 to 4-8, where figures 4-5 and 4-6 illustrate $R_S^P(V)$ and $R_S^P(VIII)$, and figures 4-7 and 4-8 illustrate $R_D^P(V)$ and $R_D^P(VIII)$. In the figures each state is represented as a node, and the arrows identify the set of infectious links predicted by the model. The red nodes represent intermediate regions, which are responsible for furthering the spread of infection. While we can numerically quantify the number of shared links (i.e. under the link cost P: $R_S^P(V)$ and $R_S^P(VIII)$ share 37% and $R_D^P(V)$ and $R_D^P(VIII)$ share 9%), these illustrations bring to light certain properties of the predictions. For example, there are certain intermediate nodes which remain constant across cases, such as Washington, California, Texas, and New York. This is likely a function of these states being infected earlier in the outbreak and the increased travel volume through the states (locations of airport hubs). The decrease in cross-country links from case V to case VIII is also apparent, resulting in shorter link distances on average. This can be attributed to the role of distance in the gravity inspired link cost for case VIII. The increased role for Texas, California and Illinois is also evident from the figures. All three of these states rank in the top five for infection count and outbound travel volume; and Texas and California were infected earliest in the outbreak so they are feasible predecessors for most other states. Overestimating the role of regions with extremely high infection reports at the end of the outbreak is a disadvantage of the static model which was previously discussed, however for case VIII the dynamic model appears to designate the same few states responsible for the majority of infection. The variability in outbreak predictions highlights the sensitivity of the model to the link weight. These illustrations also provide a way to visualize any trends in outbreak pattern behavior.

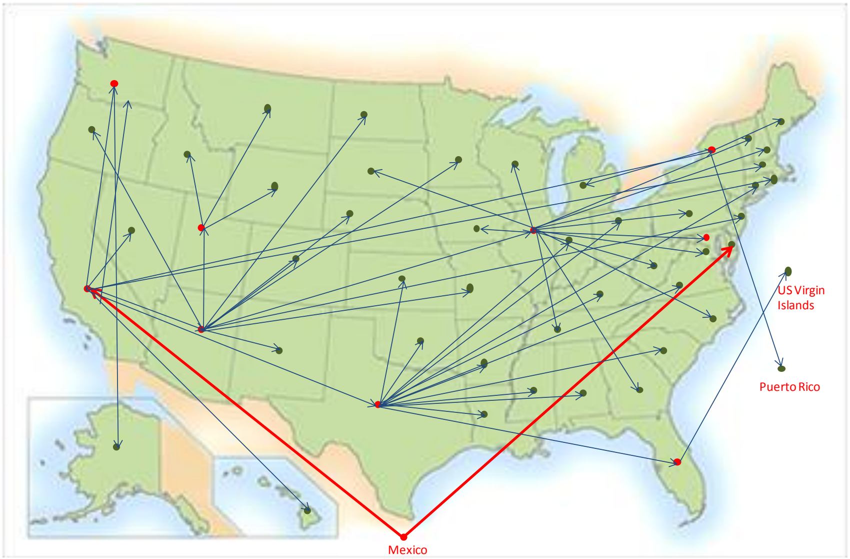


FIGURE 4-5: Mapped results for Static Case V

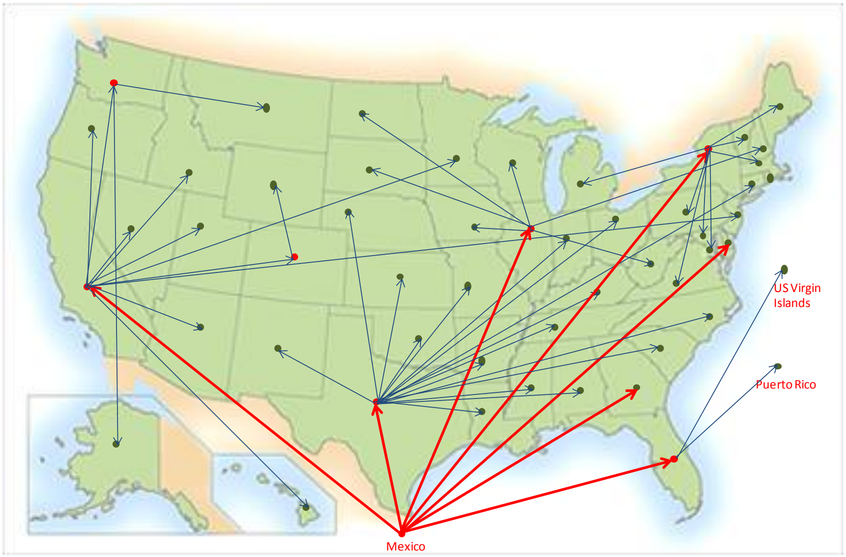


FIGURE 4-6: Mapped results for Static Case VIII

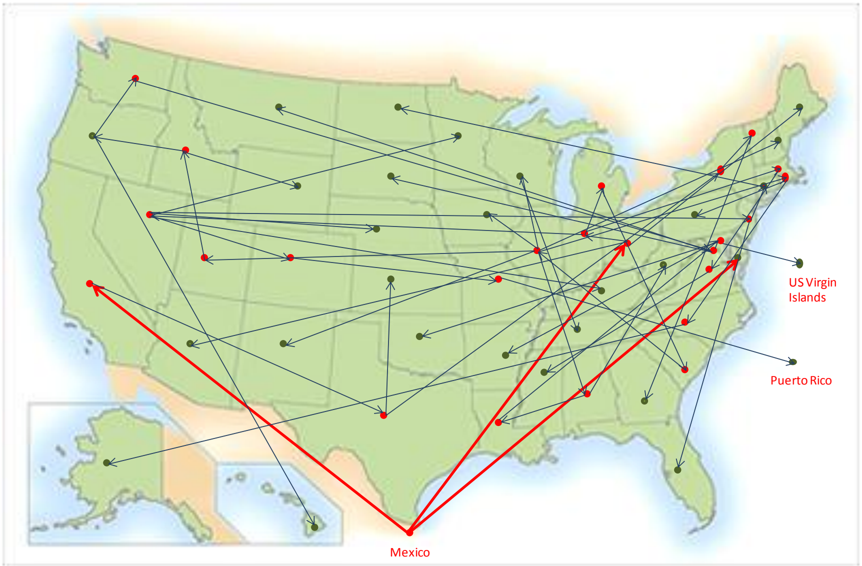


FIGURE 4-7: Mapped results for Dynamic Case V

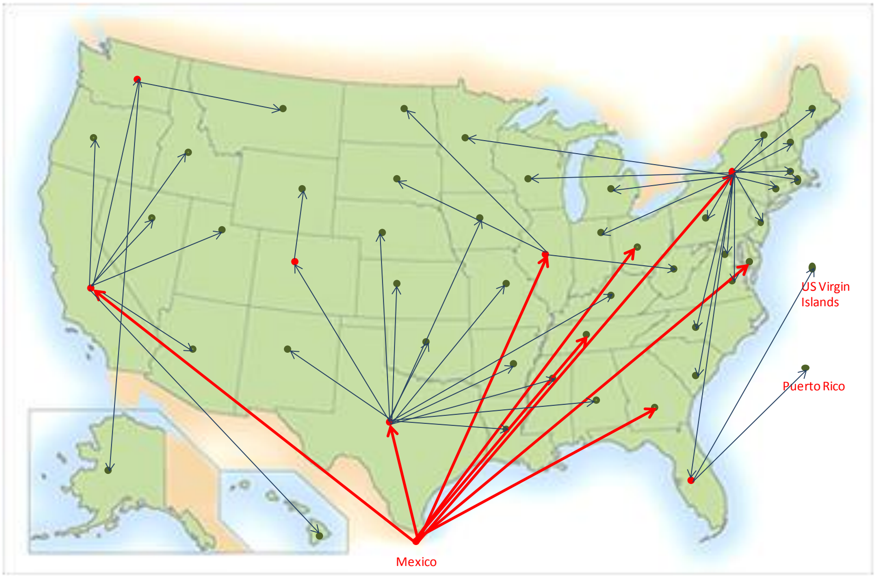


FIGURE 4-8: Mapped results for Dynamic Case VIII

4.7 CONCLUSIONS AND FUTURE RESEARCH

This chapter introduced two different modeling methodologies, i) static and ii) dynamic, to reconstruct the most likely spatiotemporal path of infection defined by human travel patterns. The proposed modeling tool is intended 1) to identify the most likely air travel routes responsible for spreading disease into new previously unexposed regions (e.g. the regions adjacent to an ongoing outbreak at highest risk), and 2) motivate regional level infection data collection efforts. Multiple link cost functions were defined and the associated outbreak scenario predictions were compared. The model predictions were also compared to a previously published phylodynamic analysis by Lemey (2009).

The robustness of the model to each variable and functional form was exposed by comparing the different case studies. With the current set of input data, neither the static nor dynamic model appears robust to variations in the link cost function and input variables. The infection data appeared to play a significant role in the model predictions, and there was little consistency between the static and dynamic models. Additionally the travel distance appears to play a larger role in the outbreak pattern than does the regional population count. The only links that were ubiquitous across models originated in Mexico, Texas, California or New York. This is likely a combined effect of the increased number of infections in these regions and the high travel volume out of these states. In addition these regions were infected earlier in the outbreak, and therefore feasible sources of infection for a longer period of time. While these variables are all intuitive factors in the spread infection to new regions, it is important to define a link weight that does not overly bias the model towards these properties.

The high level of aggregation and lack of available data made it difficult to assess model's true prediction potential. Intuitively the availability of dynamic infection data at the city level would improve dynamic model performance over the static, however at this point it is not obvious the dynamic model more accurately predicts the causal infection routes. This type of comparison requires further analysis, and more complete infection data.

In terms of comparison with Lemey's results, only seven of the 28 links listed in table 4-5 were ever predicted by the model. In general the static model predicted more of the links from Lemey (2009). Of the seven links shared, (Utah, Montana), (Florida, Tennessee) and (Mexico, Texas) were the most commonly predicted. This comparison introduces one possible application for the proposed methodology: to provide additional support for hypothesized infection spreading scenarios from similar models.

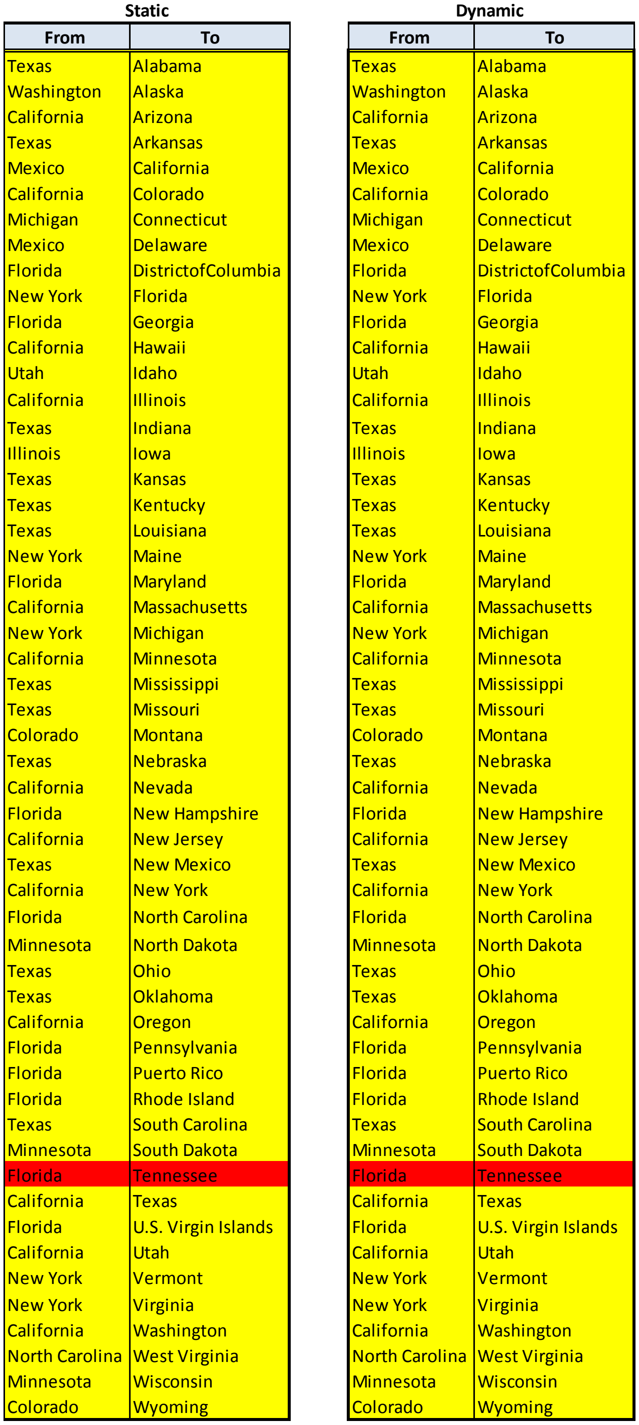
The largest weakness with the proposed methodology is the lack of verifiability due to limited data availability. Without link-based infection data to calibrate the model, it is not possible to identify the "best" prediction model, link weight or cost function. In addition the level of aggregation (to the state) results in a rather unrealistic prediction setting. The contribution of this work is more importantly a model framework, which has the potential to be expanded and applied in a much more realistic context as the necessary data becomes available. For example, one obvious extension is disaggregating the problem geographically. Predicting a city-to-city infection spreading pattern is possible with the current methodology, but implementation is solely dependent on city level infection data. Additionally, at a smaller scale many of the assumptions and model properties become more realistic.

Another potential extension of the model is to account for multiple modes of travel. The limitation imposed by assumption 4 in this work, that infections are only transmitted via air travel is highly restrictive, and likely unrealistic for many regions within close proximity. An extension of this model should have the means to include alternative modes of human transport, as well as possible freight and cargo routes capable of transporting infectious humans (or other spreading agents). A multimodal model can be defined in various ways.

One approach is to develop a *multi-layered network* where each layer represents one mode of regional transport. Under this multi-layered framework the link cost function will be mode-specific. For this approach the challenge is integrating the various levels into a single framework to define the comprehensive risk posed by a given adjacent infected region.

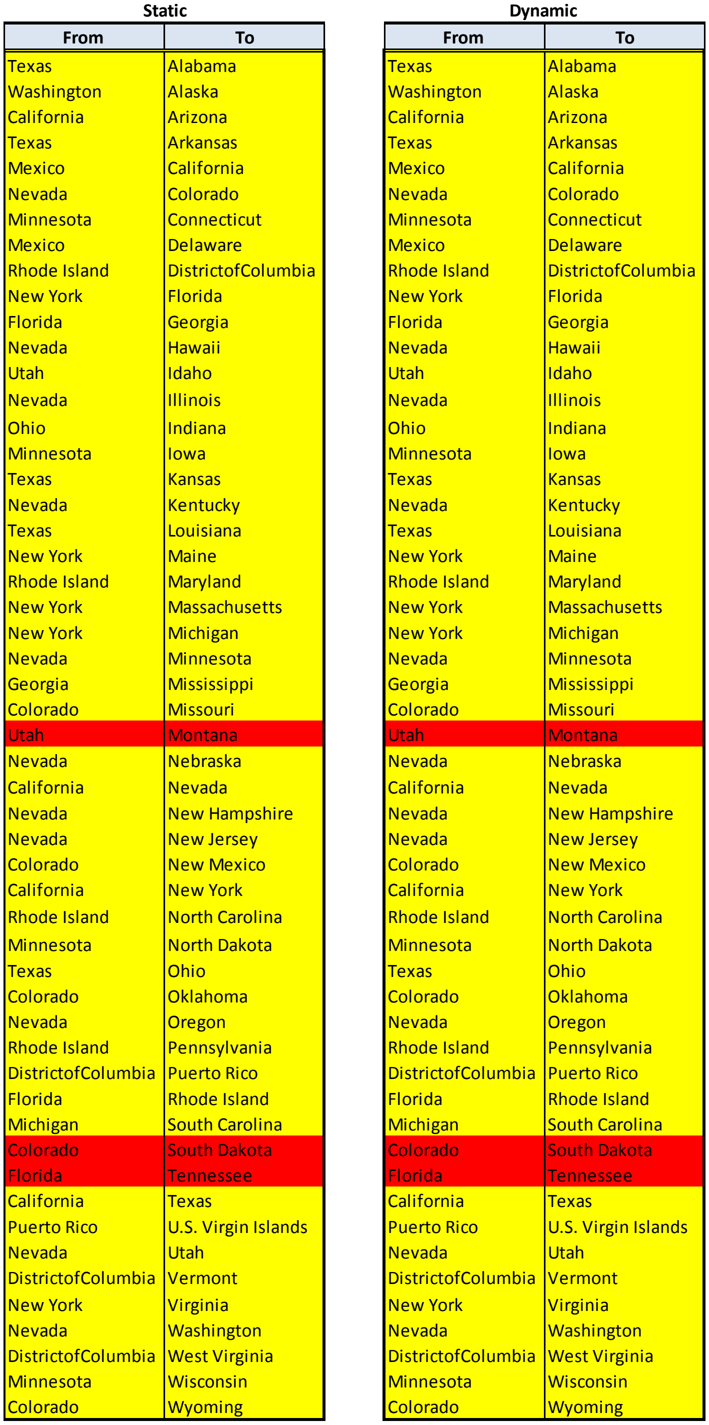
An alternative approach continues to use a *single layer network*. This is only possible if the various modes to be included can be defined using the same type of link in a network (e.g. shipping and driving cannot be aggregated into a single link volume function because they do not share travel patterns). Multi-model travel can be incorporated into the single layer network by either redefining the travel volume, v_{ij} as a new weighted function, $v_{ij}=\sum_m p_m m_{ij}$, which incorporates multiple modes of transport to represent total human mobility patterns. In this function, each travel mode, m, has a passenger volume m_{ij}, and associated mode weight $p_{m.}$ The weight allows the modes to be differentiated in terms of the likelihood of carrying an infected traveler. Another option for incorporating multi-scale human mobility takes advantage of current cellular phone tracking methods. By tracking individuals across space and time regional human mobility can be captured and quantified as a single value, v_{ij}. The methodology then remains analogous to the single mode case. These ideas will be explored in depth in future research.

TABLE 4-6: $R_S^w(I)$ and $R_D^w(I)$



Static		Dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Washington	Alaska
California	Arizona	California	Arizona
Texas	Arkansas	Texas	Arkansas
Mexico	California	Mexico	California
California	Colorado	California	Colorado
Michigan	Connecticut	Michigan	Connecticut
Mexico	Delaware	Mexico	Delaware
Florida	DistrictofColumbia	Florida	DistrictofColumbia
New York	Florida	New York	Florida
Florida	Georgia	Florida	Georgia
California	Hawaii	California	Hawaii
Utah	Idaho	Utah	Idaho
California	Illinois	California	Illinois
Texas	Indiana	Texas	Indiana
Illinois	Iowa	Illinois	Iowa
Texas	Kansas	Texas	Kansas
Texas	Kentucky	Texas	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
Florida	Maryland	Florida	Maryland
California	Massachusetts	California	Massachusetts
New York	Michigan	New York	Michigan
California	Minnesota	California	Minnesota
Texas	Mississippi	Texas	Mississippi
Texas	Missouri	Texas	Missouri
Colorado	Montana	Colorado	Montana
Texas	Nebraska	Texas	Nebraska
California	Nevada	California	Nevada
Florida	New Hampshire	Florida	New Hampshire
California	New Jersey	California	New Jersey
Texas	New Mexico	Texas	New Mexico
California	New York	California	New York
Florida	North Carolina	Florida	North Carolina
Minnesota	North Dakota	Minnesota	North Dakota
Texas	Ohio	Texas	Ohio
Texas	Oklahoma	Texas	Oklahoma
California	Oregon	California	Oregon
Florida	Pennsylvania	Florida	Pennsylvania
Florida	Puerto Rico	Florida	Puerto Rico
Florida	Rhode Island	Florida	Rhode Island
Texas	South Carolina	Texas	South Carolina
Minnesota	South Dakota	Minnesota	South Dakota
Florida	Tennessee	Florida	Tennessee
California	Texas	California	Texas
Florida	U.S. Virgin Islands	Florida	U.S. Virgin Islands
California	Utah	California	Utah
New York	Vermont	New York	Vermont
New York	Virginia	New York	Virginia
California	Washington	California	Washington
North Carolina	West Virginia	North Carolina	West Virginia
Minnesota	Wisconsin	Minnesota	Wisconsin
Colorado	Wyoming	Colorado	Wyoming

TABLE 4-7: $R_S^w(II)$ and $R_D^w(II)$



Static		Dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Washington	Alaska
California	Arizona	California	Arizona
Texas	Arkansas	Texas	Arkansas
Mexico	California	Mexico	California
Nevada	Colorado	Nevada	Colorado
Minnesota	Connecticut	Minnesota	Connecticut
Mexico	Delaware	Mexico	Delaware
Rhode Island	DistrictofColumbia	Rhode Island	DistrictofColumbia
New York	Florida	New York	Florida
Florida	Georgia	Florida	Georgia
Nevada	Hawaii	Nevada	Hawaii
Utah	Idaho	Utah	Idaho
Nevada	Illinois	Nevada	Illinois
Ohio	Indiana	Ohio	Indiana
Minnesota	Iowa	Minnesota	Iowa
Texas	Kansas	Texas	Kansas
Nevada	Kentucky	Nevada	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
Rhode Island	Maryland	Rhode Island	Maryland
New York	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
Nevada	Minnesota	Nevada	Minnesota
Georgia	Mississippi	Georgia	Mississippi
Colorado	Missouri	Colorado	Missouri
Utah	Montana	Utah	Montana
Nevada	Nebraska	Nevada	Nebraska
California	Nevada	California	Nevada
Nevada	New Hampshire	Nevada	New Hampshire
Nevada	New Jersey	Nevada	New Jersey
Colorado	New Mexico	Colorado	New Mexico
California	New York	California	New York
Rhode Island	North Carolina	Rhode Island	North Carolina
Minnesota	North Dakota	Minnesota	North Dakota
Texas	Ohio	Texas	Ohio
Colorado	Oklahoma	Colorado	Oklahoma
Nevada	Oregon	Nevada	Oregon
Rhode Island	Pennsylvania	Rhode Island	Pennsylvania
DistrictofColumbia	Puerto Rico	DistrictofColumbia	Puerto Rico
Florida	Rhode Island	Florida	Rhode Island
Michigan	South Carolina	Michigan	South Carolina
Colorado	South Dakota	Colorado	South Dakota
Florida	Tennessee	Florida	Tennessee
California	Texas	California	Texas
Puerto Rico	U.S. Virgin Islands	Puerto Rico	U.S. Virgin Islands
Nevada	Utah	Nevada	Utah
DistrictofColumbia	Vermont	DistrictofColumbia	Vermont
New York	Virginia	New York	Virginia
Nevada	Washington	Nevada	Washington
DistrictofColumbia	West Virginia	DistrictofColumbia	West Virginia
Minnesota	Wisconsin	Minnesota	Wisconsin
Colorado	Wyoming	Colorado	Wyoming

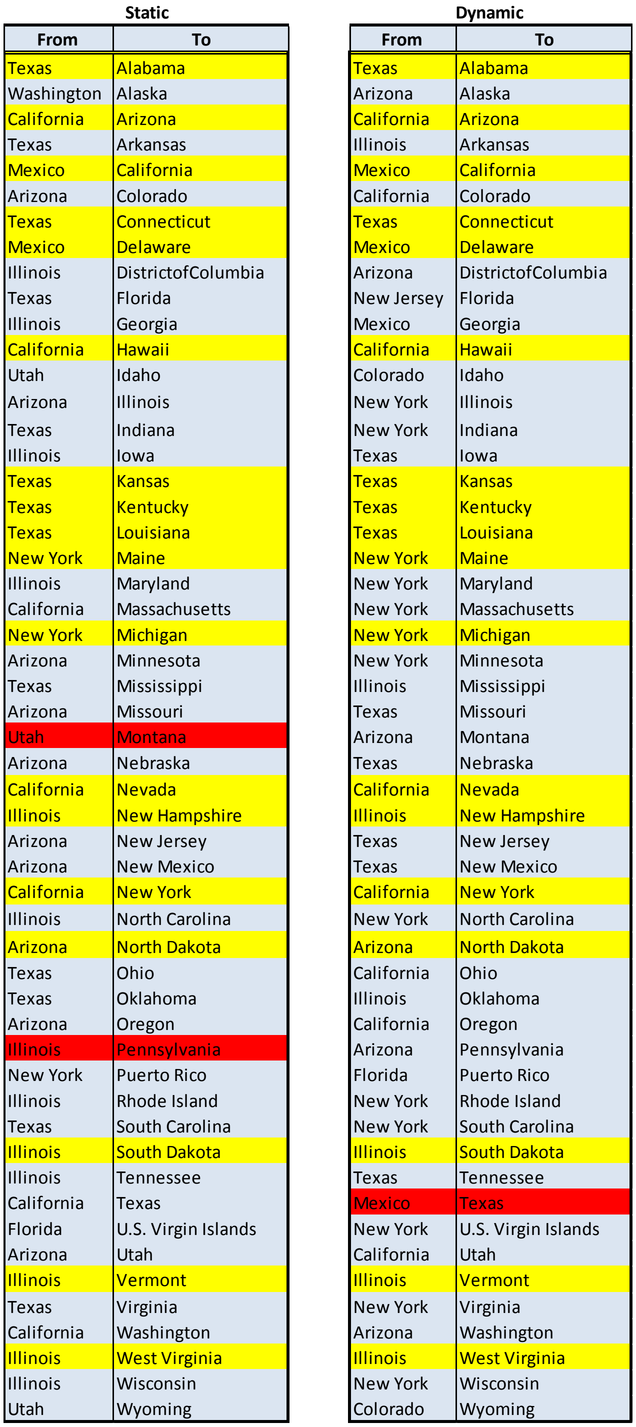
TABLE 4-8: $R_S^w(III)$ and $R_D^w(III)$

Static		Dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Washington	Alaska
California	Arizona	California	Arizona
Texas	Arkansas	Illinois	Arkansas
Mexico	California	Mexico	California
California	Colorado	Texas	Colorado
Texas	Connecticut	Texas	Connecticut
Mexico	Delaware	Mexico	Delaware
California	DistrictofColumbia	Illinois	DistrictofColumbia
Texas	Florida	New York	Florida
Texas	Georgia	Mexico	Georgia
California	Hawaii	California	Hawaii
California	Idaho	California	Idaho
California	Illinois	New York	Illinois
Texas	Indiana	New York	Indiana
Illinois	Iowa	Texas	Iowa
Texas	Kansas	Texas	Kansas
Texas	Kentucky	Texas	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
Texas	Maryland	New York	Maryland
California	Massachusetts	New York	Massachusetts
California	Michigan	New York	Michigan
California	Minnesota	New York	Minnesota
Texas	Mississippi	Texas	Mississippi
Texas	Missouri	Texas	Missouri
Washington	Montana	Washington	Montana
Texas	Nebraska	Texas	Nebraska
California	Nevada	California	Nevada
Illinois	New Hampshire	New York	New Hampshire
California	New Jersey	Texas	New Jersey
Texas	New Mexico	Texas	New Mexico
California	New York	California	New York
Texas	North Carolina	New York	North Carolina
Illinois	North Dakota	Illinois	North Dakota
Texas	Ohio	California	Ohio
Texas	Oklahoma	Texas	Oklahoma
California	Oregon	California	Oregon
Illinois	Pennsylvania	New York	Pennsylvania
New York	Puerto Rico	New York	Puerto Rico
Illinois	Rhode Island	New York	Rhode Island
Texas	South Carolina	New York	South Carolina
Illinois	South Dakota	Illinois	South Dakota
Texas	Tennessee	Texas	Tennessee
Mexico	Texas	Mexico	Texas
New York	U.S. Virgin Islands	New York	U.S. Virgin Islands
California	Utah	California	Utah
New York	Vermont	New York	Vermont
Texas	Virginia	New York	Virginia
California	Washington	California	Washington
Illinois	West Virginia	Illinois	West Virginia
Illinois	Wisconsin	New York	Wisconsin
Colorado	Wyoming	Colorado	Wyoming

TABLE 4-9: $R_S^w(IV)$ and $R_D^w(IV)$

Static		Dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Washington	Alaska
California	Arizona	California	Arizona
Texas	Arkansas	Illinois	Arkansas
Mexico	California	Mexico	California
Arizona	Colorado	Texas	Colorado
Texas	Connecticut	Texas	Connecticut
Mexico	Delaware	Mexico	Delaware
Illinois	DistrictofColumbia	Illinois	DistrictofColumbia
Texas	Florida	New York	Florida
Illinois	Georgia	New York	Georgia
California	Hawaii	California	Hawaii
Utah	Idaho	Colorado	Idaho
Arizona	Illinois	New York	Illinois
Texas	Indiana	New York	Indiana
Illinois	Iowa	Texas	Iowa
Texas	Kansas	Texas	Kansas
Texas	Kentucky	Texas	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
Illinois	Maryland	New York	Maryland
California	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
Arizona	Minnesota	New York	Minnesota
Texas	Mississippi	Texas	Mississippi
Arizona	Missouri	Texas	Missouri
Utah	Montana	Utah	Montana
Arizona	Nebraska	Texas	Nebraska
California	Nevada	New York	Nevada
Illinois	New Hampshire	Illinois	New Hampshire
Arizona	New Jersey	Texas	New Jersey
Arizona	New Mexico	Texas	New Mexico
California	New York	California	New York
Illinois	North Carolina	New York	North Carolina
Arizona	North Dakota	Minnesota	North Dakota
Texas	Ohio	California	Ohio
Texas	Oklahoma	Texas	Oklahoma
Arizona	Oregon	California	Oregon
Illinois	Pennsylvania	New York	Pennsylvania
New York	Puerto Rico	New York	Puerto Rico
Illinois	Rhode Island	New York	Rhode Island
Texas	South Carolina	New York	South Carolina
Illinois	South Dakota	Illinois	South Dakota
Illinois	Tennessee	Texas	Tennessee
California	Texas	Mexico	Texas
Florida	U.S. Virgin Islands	Florida	U.S. Virgin Islands
Arizona	Utah	California	Utah
Illinois	Vermont	New York	Vermont
Texas	Virginia	New York	Virginia
California	Washington	Illinois	Washington
Illinois	West Virginia	Illinois	West Virginia
Illinois	Wisconsin	New York	Wisconsin
Utah	Wyoming	Utah	Wyoming

TABLE 4-10: $R_S^w(V)$ and $R_D^w(V)$



Static		Dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Arizona	Alaska
California	Arizona	California	Arizona
Texas	Arkansas	Illinois	Arkansas
Mexico	California	Mexico	California
Arizona	Colorado	California	Colorado
Texas	Connecticut	Texas	Connecticut
Mexico	Delaware	Mexico	Delaware
Illinois	DistrictofColumbia	Arizona	DistrictofColumbia
Texas	Florida	New Jersey	Florida
Illinois	Georgia	Mexico	Georgia
California	Hawaii	California	Hawaii
Utah	Idaho	Colorado	Idaho
Arizona	Illinois	New York	Illinois
Texas	Indiana	New York	Indiana
Illinois	Iowa	Texas	Iowa
Texas	Kansas	Texas	Kansas
Texas	Kentucky	Texas	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
Illinois	Maryland	New York	Maryland
California	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
Arizona	Minnesota	New York	Minnesota
Texas	Mississippi	Illinois	Mississippi
Arizona	Missouri	Texas	Missouri
Utah	Montana	Arizona	Montana
Arizona	Nebraska	Texas	Nebraska
California	Nevada	California	Nevada
Illinois	New Hampshire	Illinois	New Hampshire
Arizona	New Jersey	Texas	New Jersey
Arizona	New Mexico	Texas	New Mexico
California	New York	California	New York
Illinois	North Carolina	New York	North Carolina
Arizona	North Dakota	Arizona	North Dakota
Texas	Ohio	California	Ohio
Texas	Oklahoma	Illinois	Oklahoma
Arizona	Oregon	California	Oregon
Illinois	Pennsylvania	Arizona	Pennsylvania
New York	Puerto Rico	Florida	Puerto Rico
Illinois	Rhode Island	New York	Rhode Island
Texas	South Carolina	New York	South Carolina
Illinois	South Dakota	Illinois	South Dakota
Illinois	Tennessee	Texas	Tennessee
California	Texas	Mexico	Texas
Florida	U.S. Virgin Islands	New York	U.S. Virgin Islands
Arizona	Utah	California	Utah
Illinois	Vermont	Illinois	Vermont
Texas	Virginia	New York	Virginia
California	Washington	Arizona	Washington
Illinois	West Virginia	Illinois	West Virginia
Illinois	Wisconsin	New York	Wisconsin
Utah	Wyoming	Colorado	Wyoming

TABLE 4-11: $R_S^w(VI)$ and $R_D^w(VI)$

static		dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Washington	Alaska
California	Arizona	California	Arizona
Texas	Arkansas	Texas	Arkansas
Mexico	California	Mexico	California
California	Colorado	California	Colorado
New Jersey	Connecticut	New Jersey	Connecticut
Mexico	Delaware	Mexico	Delaware
New York	DistrictofColumbia	New York	DistrictofColumbia
New York	Florida	New York	Florida
Florida	Georgia	Florida	Georgia
California	Hawaii	California	Hawaii
Utah	Idaho	Utah	Idaho
Ohio	Illinois	Ohio	Illinois
Ohio	Indiana	Ohio	Indiana
Illinois	Iowa	Illinois	Iowa
Texas	Kansas	Texas	Kansas
Ohio	Kentucky	Ohio	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
New York	Maryland	New York	Maryland
New York	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
Michigan	Minnesota	Michigan	Minnesota
Georgia	Mississippi	Georgia	Mississippi
Texas	Missouri	Texas	Missouri
Utah	Montana	Utah	Montana
Texas	Nebraska	Texas	Nebraska
California	Nevada	California	Nevada
Florida	New Hampshire	Florida	New Hampshire
Massachusetts	New Jersey	Massachusetts	New Jersey
Texas	New Mexico	Texas	New Mexico
California	New York	California	New York
Florida	North Carolina	Florida	North Carolina
Minnesota	North Dakota	Minnesota	North Dakota
Texas	Ohio	Texas	Ohio
Texas	Oklahoma	Texas	Oklahoma
California	Oregon	California	Oregon
New York	Pennsylvania	New York	Pennsylvania
Florida	Puerto Rico	Florida	Puerto Rico
Florida	Rhode Island	Florida	Rhode Island
New York	South Carolina	New York	South Carolina
Minnesota	South Dakota	Minnesota	South Dakota
Florida	Tennessee	Florida	Tennessee
California	Texas	California	Texas
Puerto Rico	U.S. Virgin Islands	Puerto Rico	U.S. Virgin Islands
California	Utah	California	Utah
New York	Vermont	New York	Vermont
New York	Virginia	New York	Virginia
California	Washington	California	Washington
DistrictofColumbia	West Virginia	DistrictofColumbia	West Virginia
Minnesota	Wisconsin	Minnesota	Wisconsin
Colorado	Wyoming	Colorado	Wyoming

TABLE 4-12: $R_S^w(VII)$ and $R_D^w(VII)$

static		dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Washington	Alaska
California	Arizona	California	Arizona
Texas	Arkansas	Illinois	Arkansas
Mexico	California	Mexico	California
California	Colorado	Texas	Colorado
New York	Connecticut	New York	Connecticut
Mexico	Delaware	Mexico	Delaware
New York	DistrictofColumbia	New York	DistrictofColumbia
Texas	Florida	New York	Florida
Florida	Georgia	New York	Georgia
California	Hawaii	California	Hawaii
California	Idaho	California	Idaho
Texas	Illinois	New York	Illinois
Texas	Indiana	New York	Indiana
Illinois	Iowa	Texas	Iowa
Texas	Kansas	Texas	Kansas
Texas	Kentucky	Texas	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
New York	Maryland	New York	Maryland
New York	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
California	Minnesota	New York	Minnesota
Texas	Mississippi	Texas	Mississippi
Texas	Missouri	Texas	Missouri
Washington	Montana	Washington	Montana
Texas	Nebraska	Texas	Nebraska
California	Nevada	California	Nevada
Illinois	New Hampshire	New York	New Hampshire
Texas	New Jersey	New York	New Jersey
Texas	New Mexico	Texas	New Mexico
California	New York	Mexico	New York
Illinois	North Carolina	New York	North Carolina
Minnesota	North Dakota	Minnesota	North Dakota
Texas	Ohio	Texas	Ohio
Texas	Oklahoma	Texas	Oklahoma
California	Oregon	California	Oregon
New York	Pennsylvania	New York	Pennsylvania
Florida	Puerto Rico	Florida	Puerto Rico
Illinois	Rhode Island	New York	Rhode Island
Texas	South Carolina	New York	South Carolina
Illinois	South Dakota	Illinois	South Dakota
Illinois	Tennessee	Texas	Tennessee
Mexico	Texas	Mexico	Texas
Florida	U.S. Virgin Islands	Florida	U.S. Virgin Islands
California	Utah	California	Utah
New York	Vermont	New York	Vermont
New York	Virginia	New York	Virginia
California	Washington	California	Washington
Illinois	West Virginia	Illinois	West Virginia
Illinois	Wisconsin	New York	Wisconsin
Colorado	Wyoming	Colorado	Wyoming

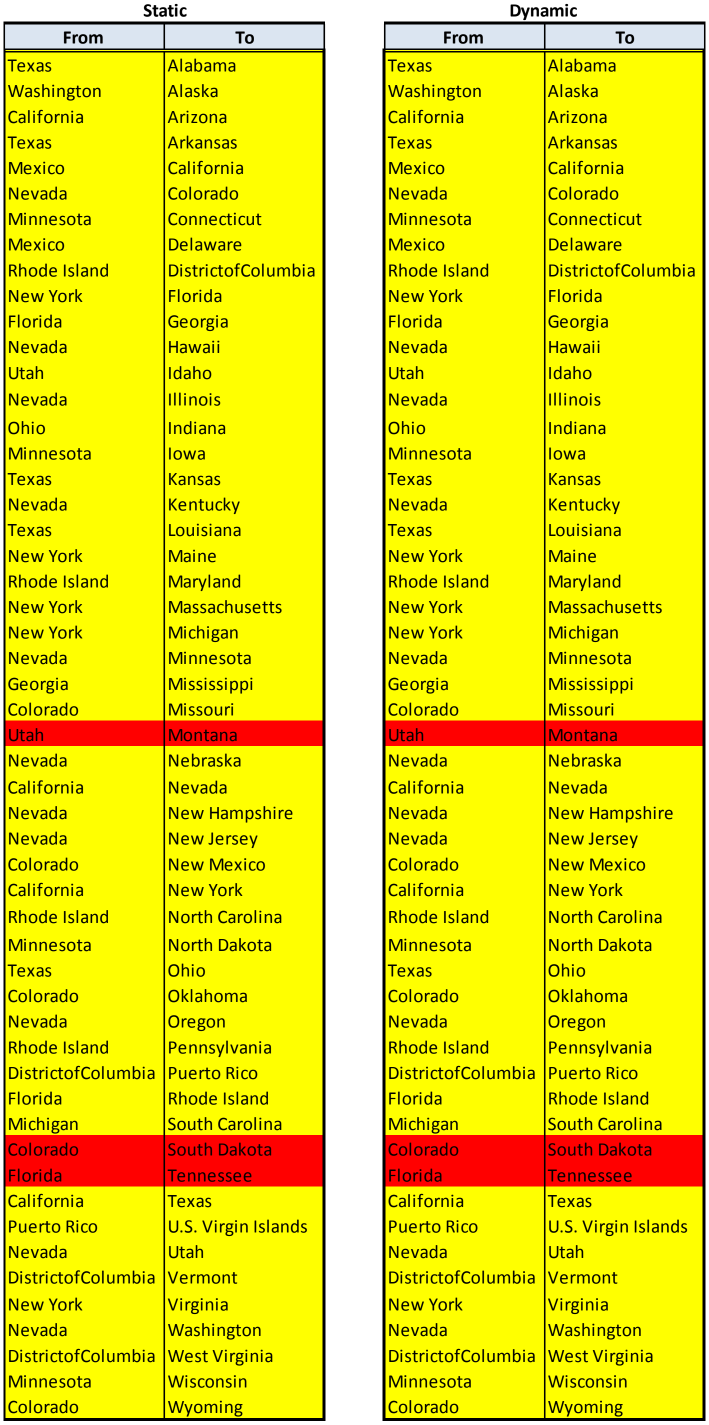
TABLE 4-13: $R_S^w(VIII)$ and $R_D^w(VIII)$

static		dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Washington	Alaska
California	Arizona	California	Arizona
Texas	Arkansas	Texas	Arkansas
Mexico	California	Mexico	California
California	Colorado	Texas	Colorado
Texas	Connecticut	New York	Connecticut
Mexico	Delaware	Mexico	Delaware
New York	DistrictofColumbia	New York	DistrictofColumbia
Mexico	Florida	New York	Florida
Mexico	Georgia	Mexico	Georgia
California	Hawaii	California	Hawaii
California	Idaho	California	Idaho
Mexico	Illinois	Mexico	Illinois
Texas	Indiana	New York	Indiana
Illinois	Iowa	Texas	Iowa
Texas	Kansas	Texas	Kansas
Texas	Kentucky	Texas	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
New York	Maryland	New York	Maryland
New York	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
California	Minnesota	New York	Minnesota
Texas	Mississippi	Texas	Mississippi
Texas	Missouri	Texas	Missouri
Washington	Montana	Washington	Montana
Texas	Nebraska	Texas	Nebraska
California	Nevada	California	Nevada
Illinois	New Hampshire	New York	New Hampshire
California	New Jersey	New York	New Jersey
Texas	New Mexico	Texas	New Mexico
Mexico	New York	Mexico	New York
Texas	North Carolina	New York	North Carolina
Illinois	North Dakota	Illinois	North Dakota
Texas	Ohio	Mexico	Ohio
Texas	Oklahoma	Texas	Oklahoma
California	Oregon	California	Oregon
New York	Pennsylvania	New York	Pennsylvania
Florida	Puerto Rico	Florida	Puerto Rico
Illinois	Rhode Island	New York	Rhode Island
Texas	South Carolina	New York	South Carolina
Illinois	South Dakota	Illinois	South Dakota
Illinois	Tennessee	Mexico	Tennessee
Mexico	Texas	Mexico	Texas
Florida	U.S. Virgin Islands	Florida	U.S. Virgin Islands
California	Utah	California	Utah
New York	Vermont	New York	Vermont
New York	Virginia	New York	Virginia
California	Washington	California	Washington
Illinois	West Virginia	Illinois	West Virginia
Illinois	Wisconsin	New York	Wisconsin
Colorado	Wyoming	Colorado	Wyoming

TABLE 4-14: $R_S^P(I)$ and $R_D^P(I)$

Static		Dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Washington	Alaska
Texas	Arizona	Texas	Arizona
Texas	Arkansas	Texas	Arkansas
Mexico	California	Mexico	California
Arizona	Colorado	Arizona	Colorado
Michigan	Connecticut	Michigan	Connecticut
Mexico	Delaware	Mexico	Delaware
Florida	DistrictofColumbia	Florida	DistrictofColumbia
New Jersey	Florida	New Jersey	Florida
Florida	Georgia	Florida	Georgia
Arizona	Hawaii	Arizona	Hawaii
Utah	Idaho	Utah	Idaho
Minnesota	Illinois	Minnesota	Illinois
Texas	Indiana	Texas	Indiana
Illinois	Iowa	Illinois	Iowa
Texas	Kansas	Texas	Kansas
Texas	Kentucky	Texas	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
Florida	Maryland	Florida	Maryland
New York	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
Michigan	Minnesota	Michigan	Minnesota
Georgia	Mississippi	Georgia	Mississippi
Colorado	Missouri	Colorado	Missouri
Utah	Montana	Utah	Montana
Texas	Nebraska	Texas	Nebraska
Texas	Nevada	Texas	Nevada
Florida	New Hampshire	Florida	New Hampshire
Texas	New Jersey	Texas	New Jersey
Texas	New Mexico	Texas	New Mexico
California	New York	California	New York
Florida	North Carolina	Florida	North Carolina
Minnesota	North Dakota	Minnesota	North Dakota
Texas	Ohio	Texas	Ohio
Texas	Oklahoma	Texas	Oklahoma
Colorado	Oregon	Colorado	Oregon
Florida	Pennsylvania	Florida	Pennsylvania
Georgia	Puerto Rico	Georgia	Puerto Rico
Florida	Rhode Island	Florida	Rhode Island
Texas	South Carolina	Texas	South Carolina
Minnesota	South Dakota	Minnesota	South Dakota
Florida	Tennessee	Florida	Tennessee
California	Texas	California	Texas
Puerto Rico	U.S. Virgin Islands	Puerto Rico	U.S. Virgin Islands
Colorado	Utah	Colorado	Utah
New York	Vermont	New York	Vermont
New York	Virginia	New York	Virginia
Colorado	Washington	Colorado	Washington
North Carolina	West Virginia	North Carolina	West Virginia
Minnesota	Wisconsin	Minnesota	Wisconsin
Colorado	Wyoming	Colorado	Wyoming

TABLE 4-15: $R_S^P(II)$ and $R_D^P(II)$



Static		Dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Washington	Alaska
California	Arizona	California	Arizona
Texas	Arkansas	Texas	Arkansas
Mexico	California	Mexico	California
Nevada	Colorado	Nevada	Colorado
Minnesota	Connecticut	Minnesota	Connecticut
Mexico	Delaware	Mexico	Delaware
Rhode Island	DistrictofColumbia	Rhode Island	DistrictofColumbia
New York	Florida	New York	Florida
Florida	Georgia	Florida	Georgia
Nevada	Hawaii	Nevada	Hawaii
Utah	Idaho	Utah	Idaho
Nevada	Illinois	Nevada	Illinois
Ohio	Indiana	Ohio	Indiana
Minnesota	Iowa	Minnesota	Iowa
Texas	Kansas	Texas	Kansas
Nevada	Kentucky	Nevada	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
Rhode Island	Maryland	Rhode Island	Maryland
New York	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
Nevada	Minnesota	Nevada	Minnesota
Georgia	Mississippi	Georgia	Mississippi
Colorado	Missouri	Colorado	Missouri
Utah	Montana	Utah	Montana
Nevada	Nebraska	Nevada	Nebraska
California	Nevada	California	Nevada
Nevada	New Hampshire	Nevada	New Hampshire
Nevada	New Jersey	Nevada	New Jersey
Colorado	New Mexico	Colorado	New Mexico
California	New York	California	New York
Rhode Island	North Carolina	Rhode Island	North Carolina
Minnesota	North Dakota	Minnesota	North Dakota
Texas	Ohio	Texas	Ohio
Colorado	Oklahoma	Colorado	Oklahoma
Nevada	Oregon	Nevada	Oregon
Rhode Island	Pennsylvania	Rhode Island	Pennsylvania
DistrictofColumbia	Puerto Rico	DistrictofColumbia	Puerto Rico
Florida	Rhode Island	Florida	Rhode Island
Michigan	South Carolina	Michigan	South Carolina
Colorado	South Dakota	Colorado	South Dakota
Florida	Tennessee	Florida	Tennessee
California	Texas	California	Texas
Puerto Rico	U.S. Virgin Islands	Puerto Rico	U.S. Virgin Islands
Nevada	Utah	Nevada	Utah
DistrictofColumbia	Vermont	DistrictofColumbia	Vermont
New York	Virginia	New York	Virginia
Nevada	Washington	Nevada	Washington
DistrictofColumbia	West Virginia	DistrictofColumbia	West Virginia
Minnesota	Wisconsin	Minnesota	Wisconsin
Colorado	Wyoming	Colorado	Wyoming

TABLE 4-16: $R_S^P(III)$ and $R_D^P(III)$

Static		Dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Oregon	Alaska
California	Arizona	New York	Arizona
Texas	Arkansas	Tennessee	Arkansas
Mexico	California	Mexico	California
California	Colorado	New York	Colorado
Texas	Connecticut	Texas	Connecticut
Mexico	Delaware	Mexico	Delaware
Illinois	DistrictofColumbia	Massachusetts	DistrictofColumbia
New York	Florida	New York	Florida
Illinois	Georgia	Florida	Georgia
California	Hawaii	Illinois	Hawaii
California	Idaho	California	Idaho
California	Illinois	Arizona	Illinois
Texas	Indiana	New York	Indiana
Illinois	Iowa	Texas	Iowa
Texas	Kansas	Texas	Kansas
Texas	Kentucky	Texas	Kentucky
Texas	Louisiana	New York	Louisiana
New York	Maine	New York	Maine
Illinois	Maryland	Florida	Maryland
California	Massachusetts	New York	Massachusetts
Texas	Michigan	New York	Michigan
California	Minnesota	New York	Minnesota
Texas	Mississippi	Georgia	Mississippi
Texas	Missouri	New York	Missouri
Washington	Montana	Washington	Montana
Texas	Nebraska	Texas	Nebraska
California	Nevada	New York	Nevada
Illinois	New Hampshire	New York	New Hampshire
California	New Jersey	New York	New Jersey
Texas	New Mexico	Texas	New Mexico
California	New York	California	New York
Illinois	North Carolina	Florida	North Carolina
Illinois	North Dakota	Illinois	North Dakota
Texas	Ohio	California	Ohio
Texas	Oklahoma	Illinois	Oklahoma
California	Oregon	Arizona	Oregon
Illinois	Pennsylvania	Arizona	Pennsylvania
New York	Puerto Rico	Pennsylvania	Puerto Rico
Illinois	Rhode Island	New York	Rhode Island
Texas	South Carolina	New York	South Carolina
Illinois	South Dakota	Illinois	South Dakota
Illinois	Tennessee	New York	Tennessee
California	Texas	California	Texas
New York	U.S. Virgin Islands	Georgia	U.S. Virgin Islands
California	Utah	New York	Utah
New York	Vermont	Illinois	Vermont
Texas	Virginia	New York	Virginia
Illinois	Washington	Illinois	Washington
Illinois	West Virginia	Florida	West Virginia
Illinois	Wisconsin	New York	Wisconsin
Colorado	Wyoming	Utah	Wyoming

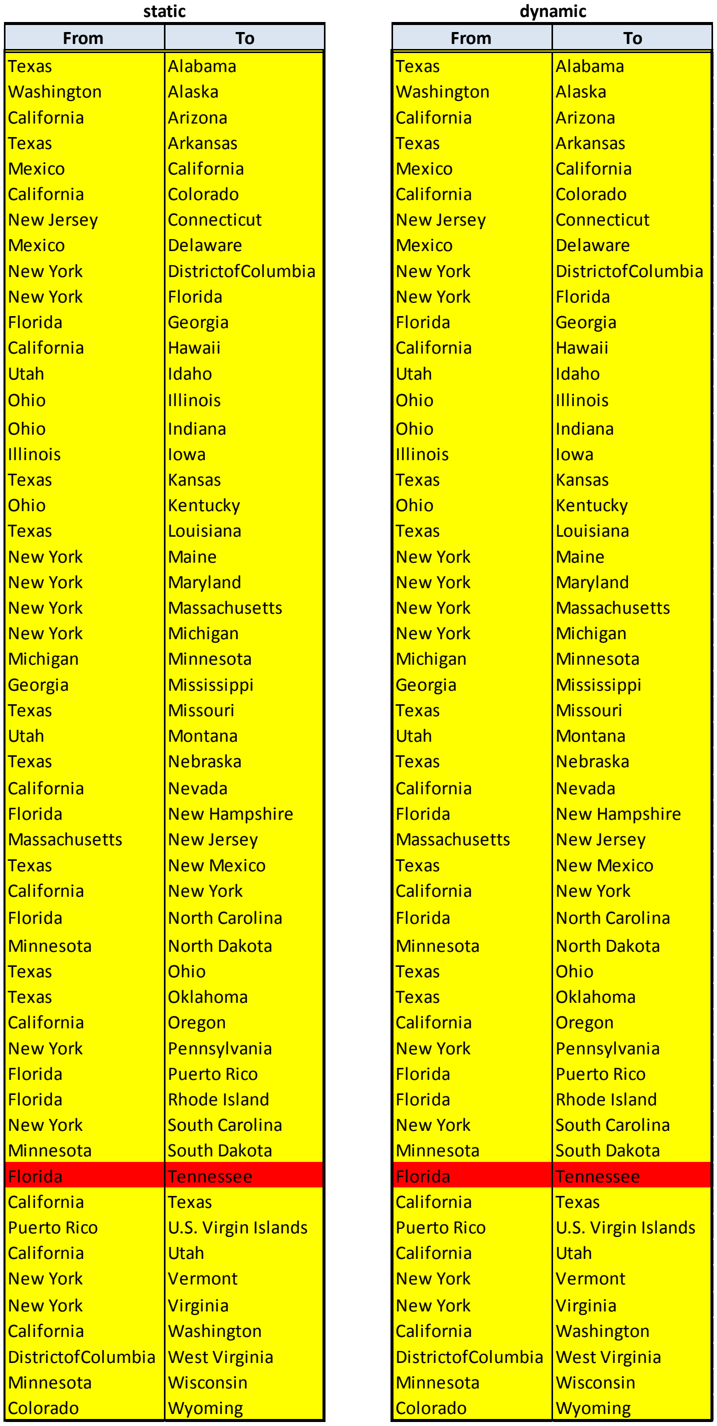
TABLE 4-17: $R_S^P(IV)$ and $R_D^P(IV)$

Static		Dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Utah	Alaska
Texas	Arizona	Mexico	Arizona
Illinois	Arkansas	Texas	Arkansas
Mexico	California	Mexico	California
Arizona	Colorado	California	Colorado
Texas	Connecticut	New York	Connecticut
Mexico	Delaware	Mexico	Delaware
Illinois	DistrictofColumbia	California	DistrictofColumbia
New York	Florida	New York	Florida
Illinois	Georgia	Texas	Georgia
Arizona	Hawaii	Colorado	Hawaii
Utah	Idaho	Colorado	Idaho
Arizona	Illinois	New York	Illinois
Texas	Indiana	Texas	Indiana
Illinois	Iowa	Texas	Iowa
Texas	Kansas	Texas	Kansas
Texas	Kentucky	Texas	Kentucky
Texas	Louisiana	New York	Louisiana
New York	Maine	New York	Maine
Illinois	Maryland	Texas	Maryland
New York	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
Arizona	Minnesota	California	Minnesota
Texas	Mississippi	Georgia	Mississippi
Arizona	Missouri	New York	Missouri
Utah	Montana	Utah	Montana
Arizona	Nebraska	Texas	Nebraska
Texas	Nevada	New York	Nevada
Illinois	New Hampshire	Illinois	New Hampshire
Arizona	New Jersey	Texas	New Jersey
Arizona	New Mexico	Texas	New Mexico
California	New York	California	New York
Illinois	North Carolina	New York	North Carolina
Arizona	North Dakota	Minnesota	North Dakota
Texas	Ohio	California	Ohio
Illinois	Oklahoma	Colorado	Oklahoma
Arizona	Oregon	California	Oregon
Illinois	Pennsylvania	New York	Pennsylvania
New York	Puerto Rico	Texas	Puerto Rico
Illinois	Rhode Island	Illinois	Rhode Island
Texas	South Carolina	Texas	South Carolina
Illinois	South Dakota	Colorado	South Dakota
Illinois	Tennessee	New York	Tennessee
California	Texas	Mexico	Texas
Florida	U.S. Virgin Islands	New York	U.S. Virgin Islands
Arizona	Utah	California	Utah
Illinois	Vermont	Illinois	Vermont
New York	Virginia	Texas	Virginia
Illinois	Washington	Mexico	Washington
DistrictofColumbia	West Virginia	DistrictofColumbia	West Virginia
Illinois	Wisconsin	Texas	Wisconsin
Utah	Wyoming	Utah	Wyoming

TABLE 4-18: $R_S^P(V)$ and $R_D^P(V)$

Static		Dynamic	
From	To	From	To
Illinois	Alabama	Illinois	Alabama
Oregon	Alaska	North Carolina	Alaska
New York	Arizona	Indiana	Arizona
Illinois	Arkansas	Maryland	Arkansas
Mexico	California	Mexico	California
Arizona	Colorado	Nevada	Colorado
Michigan	Connecticut	New Jersey	Connecticut
Mexico	Delaware	Mexico	Delaware
Washington	DistrictofColumbia	Washington	DistrictofColumbia
Michigan	Florida	New Jersey	Florida
Tennessee	Georgia	Louisiana	Georgia
Oregon	Hawaii	Oregon	Hawaii
Utah	Idaho	Utah	Idaho
Colorado	Illinois	South Carolina	Illinois
Texas	Indiana	Ohio	Indiana
Illinois	Iowa	Illinois	Iowa
Texas	Kansas	Texas	Kansas
Arizona	Kentucky	Nevada	Kentucky
Illinois	Louisiana	Alabama	Louisiana
New York	Maine	Ohio	Maine
Wisconsin	Maryland	Louisiana	Maryland
New York	Massachusetts	Indiana	Massachusetts
New York	Michigan	Indiana	Michigan
Arizona	Minnesota	Nevada	Minnesota
Illinois	Mississippi	Virginia	Mississippi
Arizona	Missouri	Colorado	Missouri
Washington	Montana	DistrictofColumbia	Montana
Arizona	Nebraska	Nevada	Nebraska
New York	Nevada	Indiana	Nevada
Illinois	New Hampshire	Illinois	New Hampshire
Arizona	New Jersey	Nevada	New Jersey
Illinois	New Mexico	Illinois	New Mexico
Texas	New York	Texas	New York
Illinois	North Carolina	Rhode Island	North Carolina
Minnesota	North Dakota	Connecticut	North Dakota
Texas	Ohio	Mexico	Ohio
Illinois	Oklahoma	Maryland	Oklahoma
Utah	Oregon	Idaho	Oregon
Wisconsin	Pennsylvania	Rhode Island	Pennsylvania
Georgia	Puerto Rico	Missouri	Puerto Rico
Illinois	Rhode Island	Illinois	Rhode Island
Michigan	South Carolina	Michigan	South Carolina
Illinois	South Dakota	DistrictofColumbia	South Dakota
Wisconsin	Tennessee	Wisconsin	Tennessee
Mexico	Texas	California	Texas
Florida	U.S. Virgin Islands	Maryland	U.S. Virgin Islands
Illinois	Utah	Illinois	Utah
Illinois	Vermont	Georgia	Vermont
New York	Virginia	Massachusetts	Virginia
Oregon	Washington	Oregon	Washington
DistrictofColumbia	West Virginia	Mississippi	West Virginia
Illinois	Wisconsin	Illinois	Wisconsin
Colorado	Wyoming	Idaho	Wyoming

TABLE 4-19: $R_S^P(VI)$ and $R_D^P(VI)$



static		dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Washington	Alaska
California	Arizona	California	Arizona
Texas	Arkansas	Texas	Arkansas
Mexico	California	Mexico	California
California	Colorado	California	Colorado
New Jersey	Connecticut	New Jersey	Connecticut
Mexico	Delaware	Mexico	Delaware
New York	DistrictofColumbia	New York	DistrictofColumbia
New York	Florida	New York	Florida
Florida	Georgia	Florida	Georgia
California	Hawaii	California	Hawaii
Utah	Idaho	Utah	Idaho
Ohio	Illinois	Ohio	Illinois
Ohio	Indiana	Ohio	Indiana
Illinois	Iowa	Illinois	Iowa
Texas	Kansas	Texas	Kansas
Ohio	Kentucky	Ohio	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
New York	Maryland	New York	Maryland
New York	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
Michigan	Minnesota	Michigan	Minnesota
Georgia	Mississippi	Georgia	Mississippi
Texas	Missouri	Texas	Missouri
Utah	Montana	Utah	Montana
Texas	Nebraska	Texas	Nebraska
California	Nevada	California	Nevada
Florida	New Hampshire	Florida	New Hampshire
Massachusetts	New Jersey	Massachusetts	New Jersey
Texas	New Mexico	Texas	New Mexico
California	New York	California	New York
Florida	North Carolina	Florida	North Carolina
Minnesota	North Dakota	Minnesota	North Dakota
Texas	Ohio	Texas	Ohio
Texas	Oklahoma	Texas	Oklahoma
California	Oregon	California	Oregon
New York	Pennsylvania	New York	Pennsylvania
Florida	Puerto Rico	Florida	Puerto Rico
Florida	Rhode Island	Florida	Rhode Island
New York	South Carolina	New York	South Carolina
Minnesota	South Dakota	Minnesota	South Dakota
Florida	Tennessee	Florida	Tennessee
California	Texas	California	Texas
Puerto Rico	U.S. Virgin Islands	Puerto Rico	U.S. Virgin Islands
California	Utah	California	Utah
New York	Vermont	New York	Vermont
New York	Virginia	New York	Virginia
California	Washington	California	Washington
DistrictofColumbia	West Virginia	DistrictofColumbia	West Virginia
Minnesota	Wisconsin	Minnesota	Wisconsin
Colorado	Wyoming	Colorado	Wyoming

TABLE 4-20: $R_S^P(VII)$ and $R_D^P(VII)$

static		dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Illinois	Alaska
California	Arizona	Texas	Arizona
Texas	Arkansas	Illinois	Arkansas
Mexico	California	Mexico	California
California	Colorado	Texas	Colorado
New York	Connecticut	New York	Connecticut
Mexico	Delaware	Mexico	Delaware
New York	DistrictofColumbia	Illinois	DistrictofColumbia
Texas	Florida	New York	Florida
Florida	Georgia	Florida	Georgia
California	Hawaii	California	Hawaii
California	Idaho	California	Idaho
Texas	Illinois	New York	Illinois
Texas	Indiana	New York	Indiana
Illinois	Iowa	Texas	Iowa
Texas	Kansas	Texas	Kansas
Texas	Kentucky	Texas	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
New York	Maryland	New York	Maryland
New York	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
California	Minnesota	New York	Minnesota
Texas	Mississippi	Georgia	Mississippi
Texas	Missouri	Texas	Missouri
Washington	Montana	Washington	Montana
Texas	Nebraska	Texas	Nebraska
California	Nevada	Texas	Nevada
Illinois	New Hampshire	New York	New Hampshire
Texas	New Jersey	New York	New Jersey
Texas	New Mexico	Texas	New Mexico
California	New York	Mexico	New York
Illinois	North Carolina	South Carolir	North Carolina
Minnesota	North Dakota	Minnesota	North Dakota
Texas	Ohio	Texas	Ohio
Texas	Oklahoma	Illinois	Oklahoma
California	Oregon	California	Oregon
New York	Pennsylvania	Texas	Pennsylvania
Florida	Puerto Rico	Illinois	Puerto Rico
Illinois	Rhode Island	New York	Rhode Island
Texas	South Carolina	New York	South Carolina
Illinois	South Dakota	Illinois	South Dakota
Illinois	Tennessee	Texas	Tennessee
Mexico	Texas	Mexico	Texas
Florida	U.S. Virgin Islands	Florida	U.S. Virgin Islands
California	Utah	California	Utah
New York	Vermont	Illinois	Vermont
New York	Virginia	New York	Virginia
California	Washington	Oregon	Washington
Illinois	West Virginia	Illinois	West Virginia
Illinois	Wisconsin	New York	Wisconsin
Colorado	Wyoming	Utah	Wyoming

TABLE 4-21: $R_S^P(VIII)$ and $R_D^P(VIII)$

static		dynamic	
From	To	From	To
Texas	Alabama	Texas	Alabama
Washington	Alaska	Washington	Alaska
California	Arizona	California	Arizona
Texas	Arkansas	Texas	Arkansas
Mexico	California	Mexico	California
California	Colorado	Texas	Colorado
Texas	Connecticut	New York	Connecticut
Mexico	Delaware	Mexico	Delaware
New York	DistrictofColumbia	New York	DistrictofColumbia
Mexico	Florida	New York	Florida
Mexico	Georgia	Mexico	Georgia
California	Hawaii	California	Hawaii
California	Idaho	California	Idaho
Mexico	Illinois	Mexico	Illinois
Texas	Indiana	New York	Indiana
Illinois	Iowa	Texas	Iowa
Texas	Kansas	Texas	Kansas
Texas	Kentucky	Texas	Kentucky
Texas	Louisiana	Texas	Louisiana
New York	Maine	New York	Maine
New York	Maryland	New York	Maryland
New York	Massachusetts	New York	Massachusetts
New York	Michigan	New York	Michigan
California	Minnesota	New York	Minnesota
Texas	Mississippi	Texas	Mississippi
Texas	Missouri	Texas	Missouri
Washington	Montana	Washington	Montana
Texas	Nebraska	Texas	Nebraska
California	Nevada	California	Nevada
Illinois	New Hampshire	New York	New Hampshire
California	New Jersey	New York	New Jersey
Texas	New Mexico	Texas	New Mexico
Mexico	New York	Mexico	New York
Texas	North Carolina	New York	North Carolina
Illinois	North Dakota	Illinois	North Dakota
Texas	Ohio	Mexico	Ohio
Texas	Oklahoma	Texas	Oklahoma
California	Oregon	California	Oregon
New York	Pennsylvania	New York	Pennsylvania
Florida	Puerto Rico	Florida	Puerto Rico
Illinois	Rhode Island	New York	Rhode Island
Texas	South Carolina	New York	South Carolina
Illinois	South Dakota	Illinois	South Dakota
Illinois	Tennessee	Mexico	Tennessee
Mexico	Texas	Mexico	Texas
Florida	U.S. Virgin Islands	Florida	U.S. Virgin Islands
California	Utah	California	Utah
New York	Vermont	New York	Vermont
New York	Virginia	New York	Virginia
California	Washington	California	Washington
Illinois	West Virginia	Illinois	West Virginia
Illinois	Wisconsin	New York	Wisconsin
Colorado	Wyoming	Colorado	Wyoming

CHAPTER 5: PREDICTING THE ROLE OF AIR TRAVEL IN SPREADING VECTOR-BORNE DISEASES

Billions of people around the globe are exposed to vector-borne diseases annually, with millions of suspected infections. Vector-borne diseases including dengue and malaria are transmitted to humans through the bite of an infected vector (i.e. mosquito). Additionally, and serving as the motivation for this research, these diseases have been increasingly reported among returning travelers in the European Union (E.U.) and United States (U.S.) (Wilder-Smith, 2005). This chapter introduces a model for quantifying the risk associated with air travel routes in the global spread of these vector-borne diseases. This model significantly varies from the previous chapters because the role of the vector in the infection process inherently alters the spreading process (compared to human contact diseases), which must be addressed.

Currently there is a lack of epidemiological surveillance on a national scale in Europe or the U.S. (Gubler, 2001). In order to limit the importation and establishment of vector-borne diseases, responsive surveillance measures must be initiated and predictive models need be developed. This analysis attempts to take a step in that direction by identifying passenger air travel routes with a high likelihood for spreading (dengue) infections into the United States and Europe from dengue-endemic regions. A network-level regression model is proposed which uses air traffic volumes, travel distances, predictive species distribution models, and infection data to quantify the likelihood of importing infection, relative to other routes. Thus, this problem has two goals:

i. To develop a model that allows planning authorities to quantify the risk from specific air travel routes, and help identify locations where local and regional surveillance systems should optimally be implemented.

ii. To highlight the importance of proper data collection efforts that should be undertaken to enhance the predictive accuracy of such models.

If provided with the necessary data, the model proposed in this chapter can be used as a prediction tool for assessing the risk of importing dengue-infected vectors or humans via air travel based on origin-destination pairs as well as to analyze the effects of changes in passenger travel routes and/or volumes on infection spreading patterns. This chapter introduces the methodology and provides a sample set of results generated using existing recent data.

The proposed methodology varies from that in the previous two chapters as follows:

i. The disease of focus is vector-borne which requires a third (non-human) spreading agent, and inherently different infection dynamics.
ii. The network structure is bipartite, where nodes fall in one of two categories:
    a. endemic regions
    b. susceptible regions
iii. Infection spreading links connect any endemic region to a susceptible one, and do not originate from a single source (the set of links identified does not result in a spanning tree).
iv. The model is *calibrated* using available infection data.

5.1 Dengue: An Emerging Disease

The application chosen for this model is Dengue fever, which has emerged as one of the most common mosquito-borne diseases in the world, the evolution of which is illustrated in Figure 5-1 (WHO, 2010). Although dengue is not currently endemic to either Europe or the continental United States, except along the Texas-México border and possibly Florida (including Key West), an increase of dengue activity in many of the endemic regions worldwide, in conjunction with a significant rise in the volume of international air travel, has resulted in a greater likelihood of imported dengue infections among travelers returning to the United States and Europe from dengue-endemic regions (Wilder-Smith, 2008). It has also increased the potential for transport and establishment of the mosquito vector species in those regions of Europe and the U.S. in which suitable habitat is available.

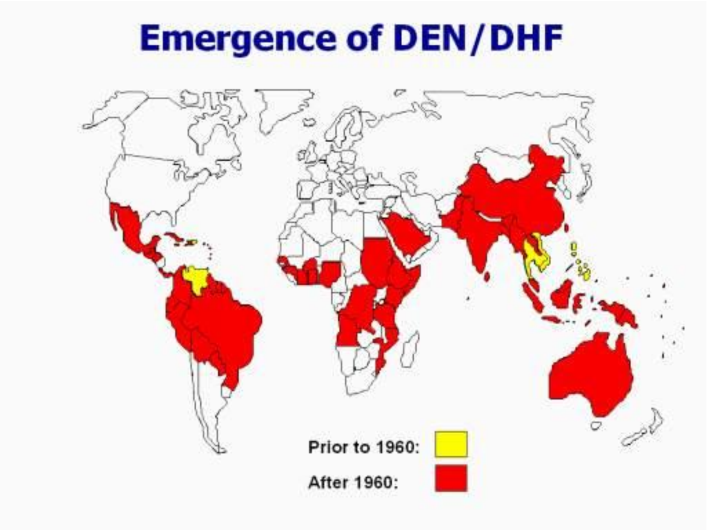


FIGURE 5-1: Map representing emergence of DF/DHF since 1960 (WHO, 2010)

Dengue viruses are transmitted from person to person through the bite of infected Aedes mosquitoes (including *Ae. aegypti* and *Ae. albopictus*), with humans serving as the main viral host (and reservoir) (WHO, 2010). The geographic establishment of dengue is thought to be limited purely by the spread of its principal vector mosquito species, *Ae. aegypti* and *Ae. albopictus*. Both species have proven to be highly adaptable to human habitation, and as a result, the global spread of the vectors can be difficult to contain (WHO, 2010). Dengue is already considered endemic to urban and suburban areas in parts of tropical and subtropical America, part of Australia, South and Southeast Asia, the Pacific, and eastern Africa. In addition, the number of imported cases of dengue in Europe and the U.S. is on the rise, and further spread and establishment in Europe and the U.S. is anticipated **(**Wilder-Smith, 2008*;* Gubler, 2001).

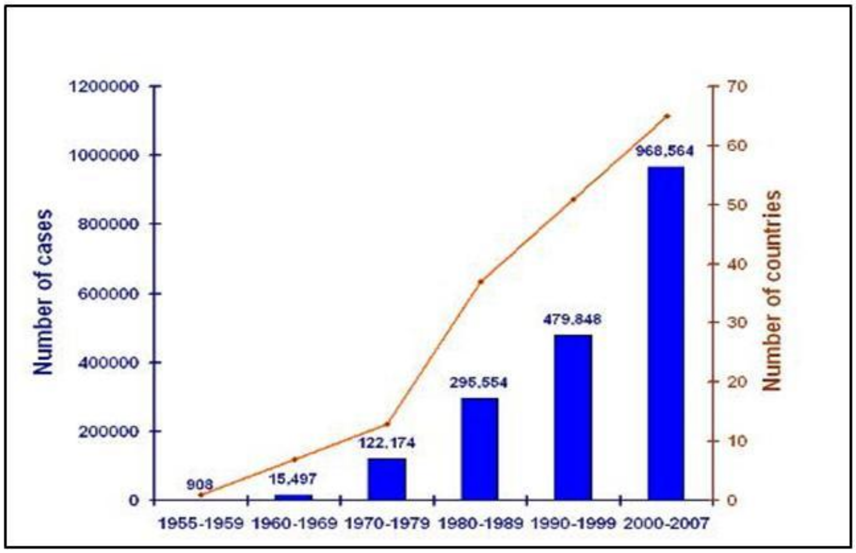


FIGURE 5-2: Graph representing the increase in reported Annual DF/DHF cases, and number of countries reporting (WHO, 2010)

Dengue infection is caused by one of four dengue virus serotypes (DENV-1, DENV-2, DENV-3, and DENV-4), ranging in clinical manifestations from asymptomatic infection to severe systemic disease (WHO, 2010). Dengue fever (DF) is the more common manifestation of the virus (an estimated 50 million infections occur annually world-wide), while dengue haemorrhagic fever (DHF) and dengue shock syndrome (DHS) are rarer and much more severe manifestations of the disease. The increasing prevalence of the disease is illustrated in Figure 5-2 (WHO, 2010). The model presented in this chapter will not distinguish between D, DHF, and DHS cases since the data available do not permit a more fine-tuned analysis.

Between 2000 and 2007 at least eight previously dengue-free areas experienced outbreaks. In 1998, an unprecedented pandemic resulted in 1.2 million cases of dengue reported from a record 56 countries worldwide, followed by a comparable situation in 2001-2002 (Wilder-Smith, 2008). Population growth, urbanization, deforestation, poor housing, inadequate sewage and waste management systems, lack of reliable water systems, and increased movement of people, pathogens, and mosquitoes contribute to continued geographic spread, increased suitability for vector species establishment, and increased incidence of the disease (Gubler, 2001). Prior to 1970, only nine countries had experienced cases of DHF; subsequently, the number has increased more than four-fold and continues to rise. Today there are at least one hundred endemic countries, with an estimated 2.5 billion people at risk. Incidence has increased 30-fold in the last 50 years. Geographic vector species' range expansion, originally promoted by sailing ships, is currently facilitated by international commercial trade (such as used tires which accumulate rain water and are a favored reproductive site, especially for *Ae. albopictus*), increased air travel, and breakdown of vector control measures (WHO, 2010).

Cyclical outbreaks of dengue fever in the U.S. remained relatively common until the early twentieth century when there were major improvements in the public health infrastructure. Although dengue causing pathogens are now rare in the U.S. and Europe, most likely due to lifestyle changes and improved living conditions (e.g., piped water systems, door and window screens, air conditioning, television), the mosquito vectors are still present. It is well-documented that at least one of the vectors capable of spreading dengue, *Ae. aegypti* or *Ae. albopictus*, has established populations in many U.S. states (Gubler, 2001). The European Center for Disease Control (ECDC, 2010) gathered entomological and environmental data to map the current distribution, as well as the risk for establishment of *Ae. albopictus* in Europe, in the event of its introduction. It concluded that temperate strains of *Ae. albopictus* currently exist and are likely to spread in several parts of Europe. In addition, new populations may become established in other parts of Europe (ECDC, 2010).

Under these conditions, imported cases of dengue via international travelers can potentially result in establishment of an autochthonous disease cycle and new regional outbreaks. This can occur in one of two ways: (i) locally established mosquito populations become infected from these new reservoirs (infected travelers) and then spread the disease; or (ii) mosquitoes carrying the virus arrive at a new environment suitable for them. This threat was exemplified recently in Key West, Florida, which experienced sizeable local outbreaks of autochthonous dengue transmission in 2009–

2010 (CDC, 2010), as well as in south Texas which has experienced dengue outbreaks in the recent past along the Texas-Tamaulipas border (CDC, 2007).

5.1.1 Role of Air Travel in Spreading Dengue

Travel is thought to be one of the leading factors in the global spread of dengue. Modern transportation bridges the natural barriers previously responsible for containing infected vectors to a specific geographic region. Today, infected humans have the potential to carry the virus into new geographical areas through air travel. A significant rise in international air traffic has increased the potential for vector dispersal into previously unoccupied regions.

Tatem *et al*'s work (discussed extensively in the literature review) served as the main motivation for this analysis, as a noticeable research gap became apparent, namely quantitative validation of such models: while the earlier work provides excellent insight into the vector importation and establishment process, the validation for the model relies on qualitative analysis of results. The proposed methodology attempts to provide further support to that approach by complementing a qualitative risk analysis with a quantitative calibration based on infection data. Further, Tatem *et al.*'s approach addresses the risk of importation and establishment of the vector and not the likelihood of infection directly. The focus of this chapter includes infected individuals, and not only the spread of disease vectors. Additionally, climatic factors are incorporated into this analysis using species distribution models, a methodology that has become standard in disease ecology and epidemiology, but was not used by Tatem and his collaborators.

5.1.2 Imported Dengue: A Threat to the United States and Europe

The proposed objective is to quantify the risk of dengue infected (air travel) passengers entering the U.S. and Europe which are regions in which dengue is not currently endemic but cases have been regularly verified. Nearly all dengue cases reported in the 48 continental United States were acquired elsewhere by travelers or immigrants. From January 1996 to the end of December 2005, 1196 cases of travel-associated dengue were reported in the continental U.S. (CDC, 2005). (Most dengue cases in U.S. nationals occur in those inhabitants of non-continental U.S. territories such as Puerto Rico [with over 5000 cases reported in 2005], the U.S. Virgin Islands, Samoa and Guam, which are all endemic regions.) In 2007, an estimated 17 million passengers traveled between the U.S. mainland and dengue-endemic areas of Asia, the Caribbean, Central and South America, and Oceania (US DOC, 2010). Since 1999 there have been 1117 cases of dengue in European travelers reported to the European Network on Imported Infectious Disease Surveillance (TropNetEurope, 2010). A recent commentary in the *Journal of American Medical Association* (JAMA) asserted that “widespread appearance of dengue in the continental United States is a real possibility” (Morens, 2008).

Further complications arise from the severe underestimation of dengue cases due to under-reporting and passive surveillance in both endemic and non-endemic regions. In tropical and subtropical countries where dengue fever is endemic, under-reporting may be due to misdiagnosis, limitations of the WHO case classification, and lack of laboratory infrastructure and resources, among other factors (Standish, 2010). To help alleviate this problem, WHO has proposed training and the adoption of standard clinical management guidelines for dengue cases. In non-endemic regions such as the United States and Europe, the actual number of dengue infections is greatly underestimated due to unfamiliarity with the disease. Additionally, 40–80% of all dengue infections are asymptomatic, and when infections are symptomatic, they often closely mimic flu symptoms. Therefore many cases may go unreported, **(**Jelinek, 2009**),** and only the most severe cases are submitted for testing. In addition, DF is currently not reported in most European public health systems. This lack of accurate infection data makes it difficult to assess the actual threat of the disease.

5.1.3 Dengue: Prevention and Control

Despite the substantial risk that dengue presents, most dengue-endemic countries have poor surveillance systems, a factor which further contributes to the spread of the disease. Various international airports implement mosquito abatement programs (such as spraying insecticides in passenger cabins); however the agent responsible for intercontinental spread of dengue infection is more likely the infected traveler, rather than infected mosquitoes (Wilder-Smith, 2008).

To address this problem, existing national public health surveillance systems should be augmented with sentinel surveillance of travelers (Wilder-Smith, 2008). Efforts currently under way include WHO's implementation of DengueNet (WHO, 2010), a data management system for the global epidemiological and virological surveillance of dengue fever (DF) and dengue haemorrhagic fever (DHF), and two (United States) Centers for Disease Control (CDC)-maintained passive surveillance systems: (ArboNET,

2010) surveillance system, a national CDC arboviral surveillance system maintained by CDC's Arboviral Diseases Branch, and the Centers for Disease Control Dengue Branch (CDCDB, 2010), a system maintained for decades which collects information on all suspected dengue cases whose specimens are sent to the branch. The model proposed in this chapter serves as a potential contribution to the development of surveillance efforts by identifying the most likely locations to encounter and interdict internationally acquired dengue infections.

5.2 PROBLEM DEFINITION

The objective is then to develop a network-based mathematical model that can be used to quantify the relative risk of importing dengue infections into the U.S. and Europe from various endemic regions around the world. The model uses passenger air traffic volumes, disease, geographic, and environmental data to determine the relative likelihood of infection being transported along a particular travel route. The proposed model predicts the expected number of dengue cases (appearing in each susceptible region) that can be attributed to each adjacent endemic region. To the best of my knowledge, the only other use of mathematical modeling to quantifying the risk estimates for acquiring dengue was proposed by Massad and Wilder-Smith (Massad, 2009). Their model is intended to evaluate the risk of infection at a specific destination as a function of human population size, the number of infected mosquitoes, and estimated parameters for the mosquitoes biting rate and the probability that an infectious mosquito will infect a susceptible human. The model does not account for travel patterns, or species distribution data; and lacks quantitative validation from infection data.

The modeling approach taken in this chapter is similar in concept to a feed-forward artificial neural network. Artificial neural networks are mathematical constructs based on the structure and workings of biological neurons. Feed-forward networks can be used to represent a learning input-output system, and can be calibrated through an algorithm called "back-propagation" to minimize a cost function which represents the output error (Bar-Yam, 1997). The approach taken in this chapter differs from traditional implementations of neural networks in that not only is a response function calibrated, but the function itself must be chosen to suit the process.

5.3 DATA

The required data for the network model include i) disease data: annual infection reports for dengue-endemic countries, susceptible European countries and susceptible U.S. states/provinces; ii) transportation data: passenger air traffic volumes for all flights originating from endemic regions and destined for Europe or the U.S; iii) geographic data: the corresponding distances for all travel routes, and iv) species distribution models which required data on the geographical occurrence of *Ae. aegypti* and *Ae. albopictus* and a suite of predictive environmental variables. The first three data sets used in this model were from 2005, and aggregated to the annual level.

The set of dengue-endemic countries is based on those identified by the CDC (CDCDB, 2010). Country level infection data for the endemic regions, as well as the European countries, were obtained from the various regional offices of the World Health Organization (WHO, 2010). U.S. state level infection data was taken from the CDC (CDC, 2005). These data sets include the annual number of reported cases for 2005 and

2007, of which the average was used to calibrate the model. The infection data is accounted for in the model in one of two ways. The number of reported cases at an endemic region is treated as an independent variable in the model; while infection reports for the susceptible node sets (U.S. states and E.U. countries) are used to calibrate the model.

There were difficulties in acquiring the necessary infection data for this model. Firstly, surveillance data for dengue in Africa are sparse. Even though all four dengue virus serotypes have been documented in the continent (Warren, under review), country-level infection data was unavailable for most African countries. These endemic countries were therefore left out of the model. Although the model would likely improve providing these data, previous research has found that Africa is responsible for the smallest percentage of travel acquired dengue infections (Rigau-Perez, 1997); thus these countries appear to be the least likely to impact the model predictions. Infection data was also unavailable for certain endemic countries in the western Pacific region. Additionally, the number of dengue cases in the U.S. and Europe are probably highly under-reported.

The transportation data was collected from two different sources. The U.S. air traffic data is from the Research and Innovative Technology Administration (RITA), a branch of the U.S. Department of Transportation (US DOT), which tracks all domestic and international flights originating or ending in the U.S. and its surrounding provinces (RITA, 2010). Passenger market data was aggregated by World Area Code (WAC) to determine the total volume of passengers traveling from each endemic country into any U.S. state in 2005. A similar analysis was done using passenger air traffic data from Eurostat (Eurostat, 2010) to determine the volume of passengers flying into each

European Union Country from each endemic country. It should be noted the transportation data used in this chapter focus on passenger travel volumes and do not include cargo flights on which vectors could potentially be transported.

The average distances used in the model were calculated in ArcGIS, an integrated Geographic Information Systems (GIS) software package. The average distances are computed for each route as the geodesic distance between the geographic centers of each region, using latitudinal and longitudinal coordinates.

5.3.1 Species Distribution Models

The risk for the establishment of dengue and potential cases of disease in an originally non-endemic area depends fundamentally on the ability of a vector to establish itself in that area. If the vector can establish itself then the disease can become endemic in two ways: (i) if the vector is already established, it can become infected from a person infected with dengue arriving in that area; or (ii) infected vectors can be transported into such an area and establish themselves. For this process, habitat in that area must be ecologically suitable for that vector. A relative measure of the suitability of one area compared to another defines a measure of the relative ecological risk (Moffett, 2007*;* González, 2010; Sarkar, 2010*;* Peterson, 2008). If the ecological risk is low, such an establishment is highly unlikely. If that risk is high, then other factors, such as the (temporally) immediate ambient environmental conditions and the size of the founder population or the availability of hosts, become critical for establishment.

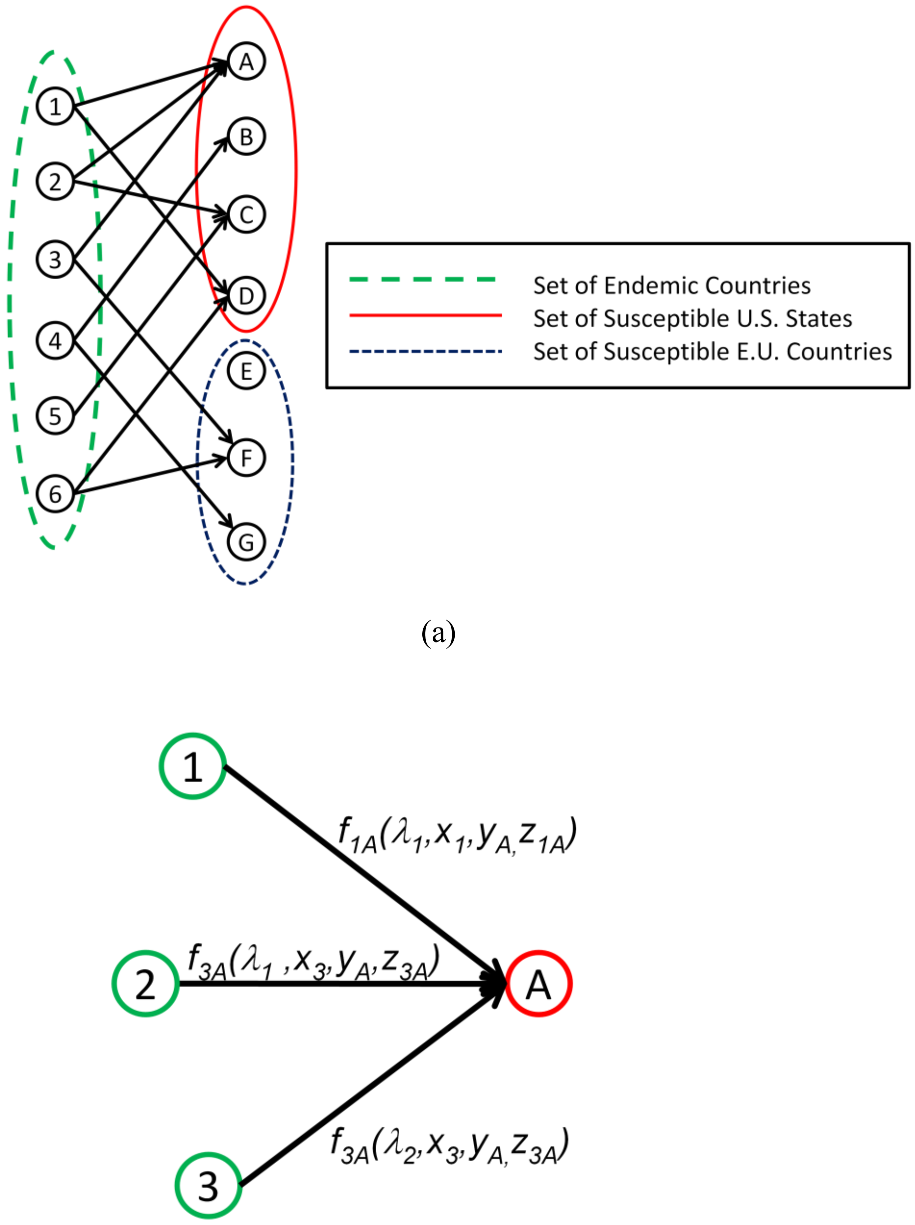
The analysis here is based on habitat suitability for the two principal dengue vector species, *Ae. aegypti* and *Ae. albopictus*. It is assumed that these two species do not interact, that is, the probability of the presence of each is independent of that of the presence of the other. The relative ecological risk for the establishment for each species is estimated using a global species distribution model at a 1 arc-minute resolution (Margules, 2007; Franklin, 2009) based on a maximum entropy algorithm incorporated in the Maxent software package (Version 3.3.4; Phillips, 2008). Maxent was used because it was predictively superior to other species distribution modeling algorithm in a large variety of studies (Franklin, 2009; Elith, 2006). As input, Maxent uses species occurrence points (presence-only data) and environmental layers (the explanatory variables). The former were obtained from the Disease Vectors database (DVD, 2010; Moffett, 2009). The latter consist of four topographic variables (elevation, aspect, slope, compound topographic index) and a standard set of 19 climatic variables all derived from the WorldClim database (Hijmans, 2005). Models were constructed using a variety of subsets of these environmental variables. All computations used default settings (Sarkar, 2010). Averages over 100 replicate models are computed. The best model was judged using the Akaike Information Criterion (AIC) for species distribution models (Warren, under review). The best model for *Ae. aegypti* is one that used all 23 explanatory variables; that for *Ae. albopictus* is based on elevation, slope, aspect, maximum temperature of warmest month, minimum temperature of coldest month, precipitation of wettest month, and precipitation of driest month. Details of the species distribution models will be published separately in the epidemiological literature.

The output from Maxent consists of relative suitability values between 0 and 1 which, when normalized, can be interpreted as the probabilistic expectation of vector presence of a species in a cell. The probabilistic expectation of at least one of the vector species being present in a cell was calculated as the complement of the probability that neither is present, assuming probabilistic independence. Because the infection and travel data used in this work are at the state level for the U.S. and the country level for Europe, the expectations are aggregated to the same level by averaging them over all the cells in the relevant geographical units. These expectations define the relative ecological risk for dengue in each cell.

5.4 Network Structure

In the proposed network structure, geographic areas are represented as nodes, belonging to either the set *G* of endemic nodes, or one of the sets N_U or N_E of susceptible nodes in the United States and Europe, respectively. The links in the network represent directed air travel connections between geographic areas (originating from *G*), while the measure P_{ji} represents the *number* of predicted infections at a susceptible node *i* attributed to an endemic node *j*.

The network structure created for this model is a directed bipartite network connecting endemic countries to susceptible regions (U.S. states and E.U. countries). Initially a single model was developed which included all susceptible regions as a single set of destination nodes, *N*. However the significantly higher number of reported infections in Europe relative to the U.S. resulted in extremely poor predictions. This is perhaps a result of unobserved variables which differentiate the risk of importing infection into Europe versus the U.S, such as border control procedures, quality of healthcare and quality of disease surveillance. Such variables are difficult to quantify directly, as found through empirical testing, and are best accounted for by using separate models. For this reason the U.S. and Europe are modeled separately for the remainder of this work. Figure 5-3.a provides an example of a bipartite network structure representative of the network structure modeled in this chapter. The network model was limited to the regions with available infection data, resulting in a network with 56 endemic nodes, 42 total susceptible nodes (30 U.S. states and 12 European countries), and 664 links. The reason the network is not fully connected is because passenger travel does not necessarily occur between all pairs of nodes.


1
2
3
4
5
6
A
B
C
D
E
F
G
Set of Endemic Countries
Set of Susceptible U.S. States
Set of Susceptible E.U. Countries
1
2
3
A
$f_{1A}(\lambda_1,x_1,y_{A,}z_{1A})$
$f_{3A}(\lambda_1 ,x_3,y_A,z_{3A})$
$f_{3A}(\lambda_2,x_3,y_{A,}z_{3A})$

(a)

(b)

FIGURE 5-3: (a) Example of bipartite network connecting endemic regions to susceptible regions, where the susceptible U.S. and Europe nodes represent mutually exclusive sets; (b) Example of link based functions to predict the number of infections at susceptible node A, attributed to each adjacent endemic region (1,2, and 3)

Figure 5-3.b is a four node extraction from the example network to illustrate the generalized link-based functional form used in our model. The function $f_{ji}(\lambda,x_j,y_i,z_{ji})$ represents the number of cases observed at i for which j is responsible, where λ represents a vector of calibrated parameters, x_j represents the characteristics of origin j, and y_i represents the characteristics of destination $i,$ and z_{ji} represents the vector of parameters specific to directed link (j,i). As such, the total predicted number of infections at i is $P_i = \sum_{\forall j \in A(i)} f_{ji}(\lambda, x_j, y_i, z_{ji})$, where $A(i)$ represents the set of endemic nodes adjacent to i.

As was the case with the problem introduced in chapter 4, the most critical issue is determining the functional form of $f_{ji}(\lambda, x_j, y_i, z_{ji})$. Two complications arise: first, the process that $f_{ji}(\lambda, x_j, y_i, z_{ji})$ attempts to model is too complex to determine a functional form *a priori*, i.e., the relative impact different variables will have is not clear ahead of time. Second, directional infection data (i.e. the source of infection for travel acquired dengue cases) are not currently available in full. Consequently, specifying the functional form of $f_{ji}(\lambda, x_j, y_i, z_{ji})$ is not feasible. In this work, the objective is to identify a link-based functional form that best replicates the number of reported cases at each susceptible region. Below is a list of the notation included in the formal problem formulation to follow.

TABLE 5-1: List of Variables

N_U	Subset of susceptible nodes in the United States
N_E	Subset of susceptible nodes in Europe
N	Complete set of susceptible nodes ($\mathrm{N_U} \cup \mathrm{N_E}$)
G	Set of nodes in the endemic region
I_i	Number of reported infections at node *i*
P_i	Total number of predicted infections at node *i*
P_{ji}	Number of predicted infections at node *i* attributed to node *j*
λ	Vector of parameter to be optimized
x_j	Vector of characteristics of infecting node *j*
y_i	Vector of characteristics of susceptible node *i*
z_{ji}	Vector if parameters specific to link (*j*,*i*)
V'_{ji}	Normalized passenger air travel volume between nodes *j* and *i,* taking a value (0,1)
S_i	Climate suitability of node *i,* taking a value (0,1)
I'_i	Normalized reported infections at node *i*
D'_{ji}	Normalized distance between nodes *j* and *i,* taking a value (0,1)
A(i)	Set of endemic nodes adjacent to susceptible node *i*
α ,β	Parameters to be optimize

5.5 SOLUTION METHODOLOGY

The purpose of this analysis then becomes to examine a variety of families of functions, further explore the most suitable member of each family, and examine the results from a qualitative perspective. The objective is to find the parameter vector λ for a given $f_{ji}(\lambda, x_j, y_i, z_{ji})$ such that the difference between I_i, the observed number of infections at susceptible node i, and P_i, the predicted number of infections at i, is as small as possible. To ensure this, a non-linear convex program is formulated to find the unknown parameter vector λ which minimizes the sum of the squared difference between observed and predicted infection values over all susceptible nodes in the set. The problem formulation is shown below:

$$\min_{\lambda} \textstyle\sum_{\forall i \in N} (I_i - P_i)^2 \qquad (5\text{-}1)$$

s.t.

$$P_{ji} = f_{ji}(\lambda, x_j, y_i, z_{ji}) \quad \forall i \in N \; \forall j \in G \qquad (5\text{-}2)$$

$$P_i = \textstyle\sum_{\forall j \in A(i)} P_{ji} \quad \forall i \in N \qquad (5\text{-}3)$$

The characteristics of the resulting linear program depend on the role of the parameter vector λ in the function, $f_{ji}(\lambda, x_j, y_i, z_{ji})$. If the function is linear in respect to λ, the resulting program can be solved analytically for the optimal decision parameters through a system of linear equations. In other cases, however, the resulting function may be non-convex, and as such solvable only through simulation.

5.5.1 Functional Forms

Depending on the functional form of $f_{ji}(\lambda, x_j, y_i, z_{ji})$, namely the behavior of $f_{ji}(\lambda, x_j, y_i, z_{ji})$ with respect to λ, the tractability of the resulting mathematical program will vary. In developing a sensible link function, several factors were considered, such as the highly nonlinear response of the explanatory variable with respect to the dependent variables considered and concerns about over fitting the data. Various functional forms were examined and compared, and the best performing function was found have the following form:

$$P_{ji} = \beta + \alpha * \frac{V'_{ji} * S_j * S_i * \sqrt{I'_j}}{\sqrt{D'_{ji}}} \quad \forall i \in N, \forall j \in G \qquad (5\text{-}4)$$

The motivation for the final functional form, P_{ji} defined above, came from the Gravity Model for Trip Distribution. The function is the sum of two terms: the first term on the R.H.S.is equivalent to the constant term in a standard regression model; while the second term bears a strong resemblance to the Gravity Model used for trip distribution. In the Gravity Model the fraction of trips attracted to zone *j* from zone *i* is proportional to the population of both zones, and inversely proportional to some measure of generalized cost of travel between them. Similarly, in the second term of the R.H.S of the equation above, the numerator accounts for the travel volume, the relative ecological risks of the origin and destination (from the species distribution models), and the number of cases reported at the source, while the distance is included in the denominator.

The square root of I'_j represents the concave relationship between the predicted number of infections at a susceptible location and the number of reported cases at an endemic source. For the denominator, the lowest value for the sum of squared errors was obtained by taking the square root of the distance. While proximity to endemic countries showed a positive correlation to the reported cases, the differential effect of distance was higher for areas closer to endemic regions. The concavity of the term can be attributed to the relationship between travel time and distance, which is certainly not linear. In order to normalize the data, the values for travel volume, distance and number of reported cases at endemic regions were rescaled by the maximum value across all observations for their respective category.

5.5.2 Model Parameter Estimation

By rewriting the original mathematical program in terms of the node based variables P_i, it is clear that it holds the same structure as a multiple linear regression, and can be solved using the Ordinary Least Squares estimation procedure:

$$\min_{\lambda} \textstyle\sum_{\forall i \in N} (I_i - P_i)^2 \qquad (5\text{-}5)$$

$$P_i = \beta * \xi(i) + \alpha * \varphi(i) \qquad \forall i \in N \qquad (5\text{-}6)$$

where:

$$\varphi(i) = \textstyle\sum_{\forall j \in A(i)} \frac{V'_{ji} * S_j * S_i * \sqrt{I'_j}}{\sqrt{D'_{ji}}} \qquad (5\text{-}7)$$

$$\xi(i) = |A(i)| \qquad (5\text{-}8)$$

In order to estimate the values of α and β, it is necessary to solve the system of equations that results from the first-order optimality conditions of the convex program shown above. The system of equations reduces to:

$$\textstyle\sum_i I_i \xi(i) - \alpha \sum_i \xi(i)\phi(i) - \beta \sum_i \xi(i)\xi(i) = 0 \qquad (5\text{-}9)$$

$$\textstyle\sum_i I_i \phi(i) - \alpha \sum_i \phi(i)\phi(i) - \beta \sum_i \phi(i)\xi(i) = 0 \qquad (5\text{-}10)$$

Solving the system of equations yields the following estimates for α and β:

$$\alpha = \frac{\sum_i I_i \xi(i) \sum_i \xi(i)\phi(i) - \sum_i I_i \phi(i) \sum_i \xi(i)\xi(i)}{\sum_i \xi(i)\phi(i) \sum_i \xi(i)\phi(i) - \sum_i \xi(i)\xi(i) \sum_i \phi(i)\phi(i)} \qquad (5\text{-}11)$$

$$\beta = \frac{\alpha \sum_i \phi(i)\phi(i) - \sum_i I_i \phi(i)}{\sum_i \xi(i)\phi(i)} \qquad (5\text{-}12)$$

5.6 Numerical Results and Analysis

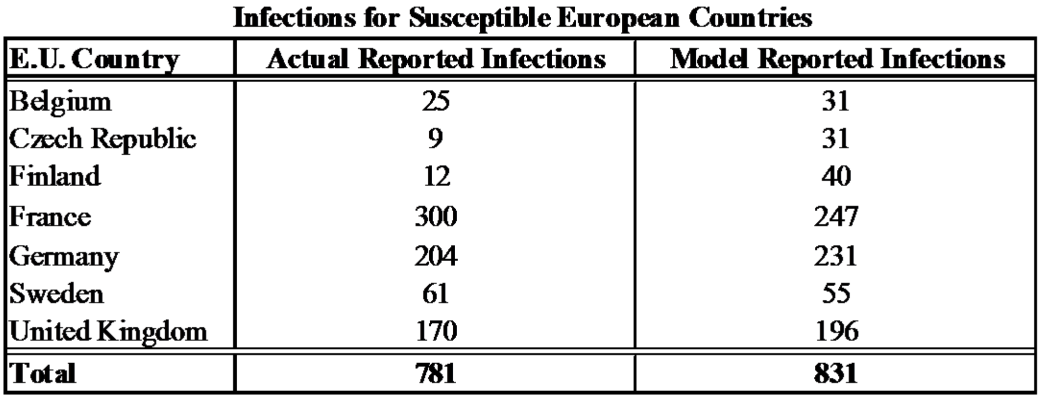
The main objective of the model is to quantify the relative risk of various international travel routes. This is accomplished by first predicting the number of dengue cases specific to each travel route, and then calibrating the network model at a regional level using infection data. Therefore, there are two sets of results presented here. Section 5.5.1 includes the total number of dengue cases predicted for each susceptible region based on the calibrated model output, and Section 5.5.2 include the corresponding relative risk of each travel route, ranked based on their likelihood of transporting infected passengers.

The results included in this section are representative of filtered data. The filtering process is applied to the susceptible node set to remove outliers. The outliers are classified differently for the European and U.S. node sets. In the European data set any region with less than 5 cases was considered an outlier, while only states with one reported case are considered outliers in the U.S. node set. A lower threshold was implemented for the U.S. as there were fewer reported cases on average. The procedure resulted in five nodes being removed from N_E and 12 nodes being removed from N_U. After the filtering process there were 18 U.S. states and seven European countries included in the model.

5.6.1 Susceptible Node-Based Predictions

The model was able to predict closely the number of reported cases for the European countries, though it struggled to predict the number of reported cases for the U.S. states accurately. The results for the node-based predictions, P_i, are shown in Table 5-2.a for European Countries and Table 5-2.b for U.S. states

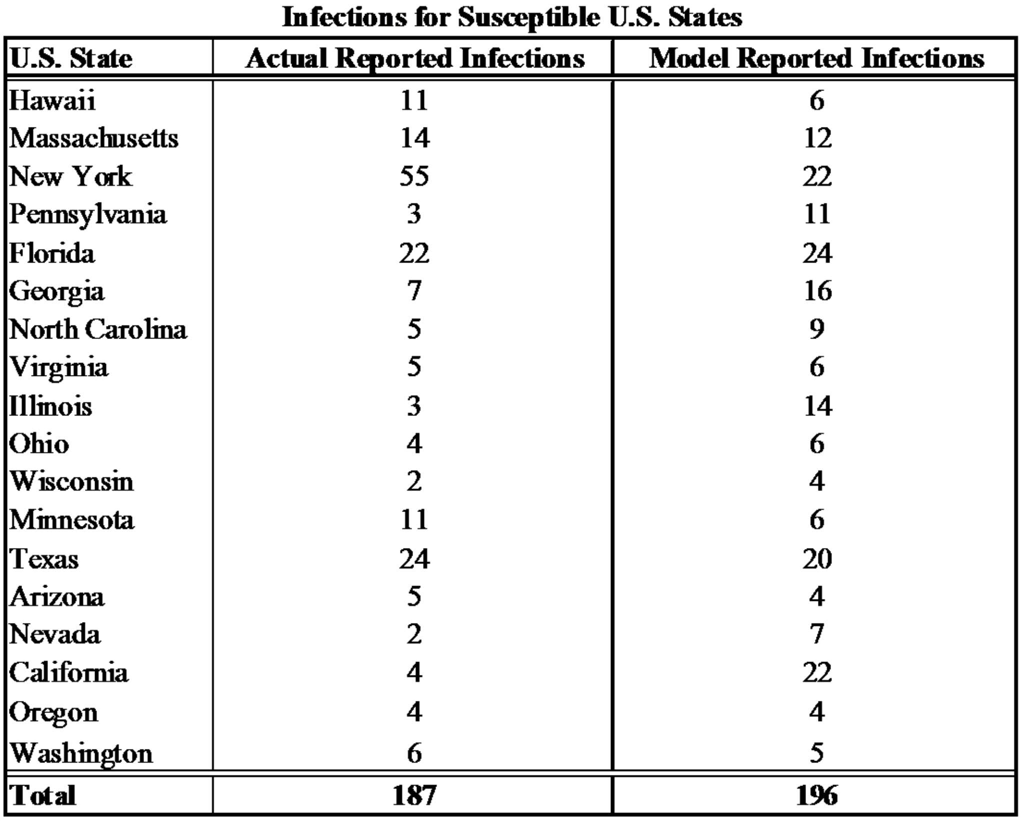
TABLE 5-2: Model output and actual reported infections for (a) Europe and (b) U.S.



Infections for Susceptible European Countries

E.U. Country	Actual Reported Infections	Model Reported Infections
Belgium	25	31
Czech Republic	9	31
Finland	12	40
France	300	247
Germany	204	231
Sweden	61	55
United Kingdom	170	196
Total	**781**	**831**

(a)



Infections for Susceptible U.S. States

U.S. State	Actual Reported Infections	Model Reported Infections
Hawaii	11	6
Massachusetts	14	12
New York	55	22
Pennsylvania	3	11
Florida	22	24
Georgia	7	16
North Carolina	5	9
Virginia	5	6
Illinois	3	14
Ohio	4	6
Wisconsin	2	4
Minnesota	11	6
Texas	24	20
Arizona	5	4
Nevada	2	7
California	4	22
Oregon	4	4
Washington	6	5
Total	**187**	**196**

(b)

The same functional form introduced in section 5.4.3 was used in both models, while the resulting regression parameters, α and β were highly variable. For Europe the optimal α and β were 271.52 and 5.08 respectively; for the U.S. 5.54 and 0.595. The combination of the low constant (β), high α value, and good fit of the European model signifies that the majority of variability in the data was accounted for by the independent variables included in the model. This was not the case with the U.S. model. On average, the European model predictions diverged from the reported cases by 24, where 112 actual cases were observed on average per node. The U.S. model predictions diverged from the reported cases by an average of 6.2, where an average of 10.4 cases were reported per node.

Several factors contribute to complicating the task of identifying a function to perfectly fit the case data. Firstly, the limited size of the susceptible node set makes it difficult for the model to differentiate between variability and noise. Secondly, the amount of noise in the data due to unknown factors such as variations in regional surveillance efforts cannot be accounted for. Thirdly, current prevention measures being implemented are not only difficult to determine, but also difficult to quantify. All these uncertainties restrict the model's ability to estimate parameters that result in good predictive properties at the node level. However, our results show that, while the fit at the node level could be improved upon, the route-level risk measures do show promising results, and as such, provide some insight into the role the independent variables play.

5.6.2 Endemic-Susceptible Route-Based Risk

Although the node-based predictions can be validated based on the reported infection data, the resulting route-based predictions are not directly-verifiable due to the unavailability of route-based infection data. The best measures of validation are (i) to find route-based predictions that correspond to known regional infection data when summed across all incoming routes, and (ii) to compare the results with previous travel-based patient surveys conducted to determine the most likely place of origin for illness.

Table 5-3 identifies the 20 international travel routes with the highest probability of carrying dengue infected passengers into (a) Europe and (b) the U.S., and their corresponding relative risk, as produced by the model. The initial ranking was determined based on the predicted number of infected passengers traveling on each route. The predicted number of infected passengers was then normalized to the highest ranked route. Although the results shown are specific to the filtered node sets, similar results were obtained for the full node sets, for both Europe and the U.S. In the model Burma, Cambodia, Laos, and Thailand are aggregated to a single "South East Asia" endemic region.

TABLE 5-3: Relative risk of spreading travel acquired dengue infection via international travel routes from endemic countries into (a) Europe and (b) U.S.

Route-Based Relative Risk for European Countries

Rank	From	To	Relative Risk
1	Brazil	Germany	1.00
2	Brazil	France	0.99
3	South East Asia	Germany	0.71
4	South East Asia	United Kingdom	0.52
5	Brazil	United Kingdom	0.35
6	South East Asia	France	0.29
7	Vietnam	France	0.29
8	Singapore	United Kingdom	0.27
9	Singapore	Germany	0.19
10	India	Germany	0.19
11	Malaysia	United Kingdom	0.19
12	India	United Kingdom	0.17
13	Dominican Republic	Germany	0.16
14	Venezuela	Germany	0.16
15	Dominican Republic	France	0.16
16	Mexico	France	0.16
17	Mexico	Germany	0.15
18	Venezuela	France	0.15
19	South East Asia	Finland	0.14
20	South East Asia	Sweden	0.13

(a)

TABLE 5-3, continued

Route-Based Relative Risk for U.S. States

Rank	From	To	Relative Risk
1	Mexico	Texas	1.00
2	Mexico	California	0.56
3	Puerto Rico	Florida	0.34
4	Brazil	Florida	0.33
5	Venezuela	Florida	0.24
6	Mexico	Illinois	0.23
7	Puerto Rico	New York	0.21
8	Costa Rica	Florida	0.19
9	Mexico	Florida	0.19
10	Mexico	Arizona	0.19
11	Dominican Republic	New York	0.17
12	Colombia	Florida	0.16
13	Brazil	New York	0.15
14	Mexico	Georgia	0.15
15	Dominican Republic	Florida	0.15
16	Brazil	Texas	0.14
17	Brazil	Georgia	0.12
18	Honduras	Florida	0.12
19	Costa Rica	Texas	0.12
20	Mexico	Nevada	0.11

(b)

As stated previously, one way of verifying the predicted route-based risk is by comparing the results with previous patient surveys conducted to identify the source of infections. A previous study found of the travel acquired dengue cases in Europe between 1999 -2002 (Wichmann, 2003):

i. 219 (45%) originated in South-East Asia, represented in the model as 3 of the top 6 highest risk routes.

ii. 91 cases (19%) originated in South and Central America, represented in the model as 3 of the top 10 highest risk routes.

iii. 77 cases (16%) originated in the Indian subcontinent, represented in the model as 2 of the top 15 highest risk routes.

iv. 56 cases (12%) originated in the Caribbean, represented in the model as 2 of the top 20 highest risk routes.

The model predicts Brazil-Germany and Brazil-France as the two highest risk routes into Europe (with nearly equivalent relative risk). This is expected, as Brazil reports the highest number of dengue cases in the world per year, almost 3 times those of second place Indonesia, and the volume of traffic on the Brazil-France and Brazil-Germany routes are two of the top 40 in the world. Indonesia, while reporting a very high number of infections, reports very low levels of air travel on any given route destined for Europe. Using similar logic, Southeast Asia reports a number of infections on par with Indonesia, though the travel volume from Southeast Asia into Germany and the United Kingdom rank among the world's top 25 travel routes; suggesting intuitively that travel volume is a dominant factor in assessing infection risk.

For the U.S. the model predicts the majority of U.S. infections are attributed to Central and South American countries, likely a result of the close proximity, high traffic, and high level of infection. More specifically, 19 of the top 20 highest risk routes into the US (Nevada, ranked 20th not included) are destined for states which account for a very high fraction of incoming flights in the US; accounting for 6 of the top 15 busiest American Airports by boardings (FAA, 2010).

As a destination, Florida accounted for 5 of the top 10 risk routes, which is supported by historical occurrence of the disease, as exemplified in the 2009–2010 local outbreaks. Though it is possible that dengue was already present in the locality (Key West), and previously undetected, the results of this model suggest dengue could likely have been introduced via international travelers into a locality with environmental and social conditions ripe for transmission (CDC, 2010). This is represented in the model as Puerto Rico-Florida ranks as the third highest risk route. This travel volume on this route is among the top ten in the world, while the proximity and climate similarity are also likely contributors to the infection risk.

Mexico-Texas and Mexico-California rank as the two highest risk routes, and are also the top two traveled routes (by passenger volume) in the world (RITA, 2010). The highest risk route predicted for infection is between Mexico and Texas; nearly twice that of Mexico-California; which is also supported by historical data with outbreaks reported as recently as 2005 in Brownsville (CDC, 2007). These local outbreaks were attributed to concurrent outbreaks in neighboring Mexican border towns. The high number of infections reported in Mexico, its proximity to Texas, and the high volume of travel between the two intuitively suggests this to be a high risk pairing, which is supported by the model.

5.7 Conclusions

Today, dengue poses a serious threat to many parts of the U.S. and Europe where suitable environmental conditions for vector species provide the potential for local outbreaks, were the disease to be introduced, besides the potential for new vector species' population establishment. This work was motivated by the increasing number of dengue diagnoses in the U.S. and Europe, coinciding with an increase in both the prevalence of dengue worldwide and increased volume of international passenger air traffic originating from dengue endemic regions.

The model was developed as a means to explore the relationship between reported dengue infections and air travel. The model implements a network-based regression methodology to quantify the relative risk from international air travel routes carrying passengers from dengue endemic regions to susceptible regions in the U.S. and Europe. In addition to international passenger travel volumes, the model takes into account predictive species distribution models for the principal vector mosquito species. The model also incorporates travel distances and infection data. The following inferences can be drawn from the model results:

i. The highest risk travel routes suggest that the proximity to endemic regions is a dominant factor. Most high risk routes into Europe originate in Asia (with the exception of Brazil and Mexico), while all top 20 routes into the U.S. originate in South and Central America.

ii. Travel from dengue-endemic countries poses a significant threat for Florida. Additionally, the high volume of domestic visitors to Florida in conjunction with an established *Ae. albopictus* population, provides an additional complexity to Florida's role as a gateway for dengue into other parts of the U.S. The recent reemergence of dengue in Florida suggests strong vector-borne surveillance and mosquito control infrastructure will be crucial for identification and control of outbreaks of dengue.

iii. The high risk predicted for Mexico-Texas travel is further heightened by the risk of overland transmission (such as that from Tamaulipas into the Brownsville area (*8*)). Therefore surveillance along the Texas-Tamaulipas border should be complimented with surveillance at regions with airports connected to Mexico by regular or chartered flights.

iv. For many regions of Europe and the U.S., if dengue gets introduced, the establishment of an autochthonous disease cycle is likely because many of these areas contain suitable habitats for *Ae. albopictus*.

v. Some of the "source" areas indicate that dengue has yet to be brought under control in places where malaria has. This means that dengue may well replace malaria as the paradigmatic airport disease.

The results provided in this chapter were determined using existing (historical) data from the (recent) past, and do not represent accurate predictions for the relative risks for future scenarios. However the objective of this chapter is to introduce a model (which can be further improved given a more complete set of data), that can be calibrated using epidemiological data. The calibrated model can be used as a predictive tool for quantifying route-based risk if provided with the necessary data including real-time travel patterns, environmental conditions, infection data, etc. In addition the results in this chapter are aggregated to the annual and regional (country, province or state) level, due to the available data.

The development of such a model is an integral step in improving local and regional surveillance efforts. The quantitative results produced by the model can lead to more specific surveillance recommendations than the CDC is able to make, such as identifying specific routes on which to implement control strategies, and identifying locations (origin cities, destination airports, etc.) at which passenger surveillance efforts would be most beneficial. As there is currently no vaccine for dengue; surveillance and intervention, along with vector control, are the leading options in preventing further geographic distribution. This research also highlights the need for improved quality in disease data, and how these data can help better predict and control epidemic episodes of vector-borne diseases in susceptible countries.

5.8 Future Research

The methodology introduced in this chapter has substantial room for improvement. For the application presented infection data was the limiting variable, for example infection reports for many regions in the world are not available even at the annual level. More complete infection data would allow for more advanced analysis. Direct extensions of this model include i) regional disaggregation (i.e. to the city level) and ii) temporal disaggregation to account for seasonality. The proposed methodology can also be directly applied to alternative applications such as iii) geographical regions, iv) modes of transportation and v) vector-borne diseases.

5.8.1 Potential Extension to link-based formulation

The extensions listed above do not require significant changes to the solution methodology or network structure. One potential extension of the model requiring significant methodological innovation replaces the single-link routes with multi-link travel paths. In the current model airport layovers are ignored; travel routes are direct links between endemic and susceptible regions. This makes the implicit assumption that infections are not transmitted at airports during layovers, which is likely unrealistic. To address this issue the bipartite network structure would have to be relaxed, and replaced with a traditional air traffic network structure. The current path-level predictions would have to be replaced with link-level predictions. In addition to the complexity introduced by expanding the network structure, other challenges are introduced. For example a route from Thailand to New York might have a stopover in London, at which point an infected passenger can spread infection to another individual in the airport. This type of occurrence is difficult to track, let alone predict, because the individual infected at the London airport might be local, or en route to some other destination. From this example, it is obvious there are no longer mutually exclusive sets differentiating endemic and susceptible regions. To solve this network level problem a new link-based prediction function would need to be developed, and some sort of node balance constraints which track the infections along a travel path must be introduced. This is a problem which will be addressed in future research.

5.8.2 Potential Extension to multi-mode network

Another extension that requires innovative solution methodologies incorporates multiple modes of transportation, such as freight and shipping networks, into a single integrated model. While the proposed methodology can be directly applied to alternative network structures which characterize a single mode of transportation (respective travel routes and volumes); integrating multiple modes into a single model poses new challenges. For example, if there are multiple incoming modes of transport for a single region, which mode is responsible for the imported infections? Or how should the responsibility be split between modes?

Proposed Problem Definition

Modeling the dispersal of dengue across geographic space serves as an ideal application for a multi-layered network problem. As an example, two types of interaction networks include: (i) the passenger air travel network analyzed in this chapter; and (ii) the shipping cargo network with ports as vertices and shipping routes as links. These two systems can naturally be layered, and represented as a single integrated network model.

Using respective properties of each transportation network structure and the additional data sets such as in the single mode model, an analogous integrated network-based mathematical model is sought that can be used to quantify the relative risk of importing dengue infections into susceptible regions from various endemic regions around the world.

Problem Description

To model the integrated transportation network, since both air ports and (maritime) ports are spatial locations, a single set of vertices may be connected by two different types of links, one corresponding to air travel routes and the other corresponding to cargo transport routes. The network structure created for this model is still a directed bipartite network connecting endemic countries to susceptible regions. The geographic areas are represented as nodes, belonging to either the set *G* of endemic nodes, or the set *N* of susceptible nodes, respectively. The links in the network represent directed mode-specific (k) travel connections between geographic areas (originating from *G*), weighted by the volume of passenger or cargo usage. The link associated measure P_{ji}^{k} represents the number of predicted infections at a susceptible node *i* attributed to an endemic node *j* specific to mode *k*.

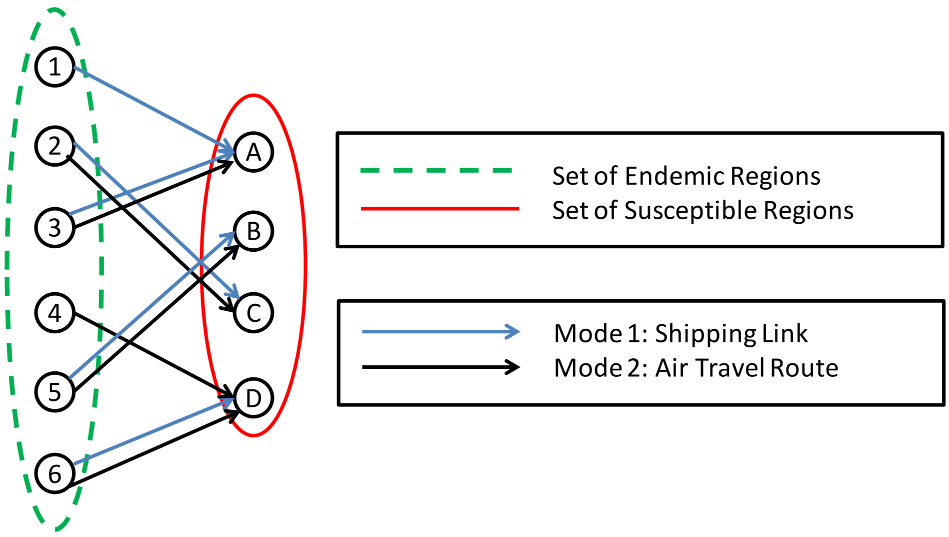


FIGURE 5-4: Example of multimodal bipartite network connecting endemic regions to susceptible regions

Modeling Methodology

The objective of the model is to quantify both the risk associated with passenger travel routes in terms of infected individuals, and the risk associated with cargo routes in terms of infected vectors. The objective is therefore to define a link based function specific to travel mode k, $f_{ji}^{k}(\lambda_k, x_j, y_i, z_{ji}^{k})$ to predict the risk of importing either infected humans or vectors at each susceptible node i, attributed to each adjacent endemic region j, where λ_k represents a vector of calibrated parameters for mode k, x_j represents the characteristics of origin j, y_i represents the characteristics of destination i, and z_{ji}^{k} represents the vector of characteristics specific to directed link *(j,i)* and mode *k*.

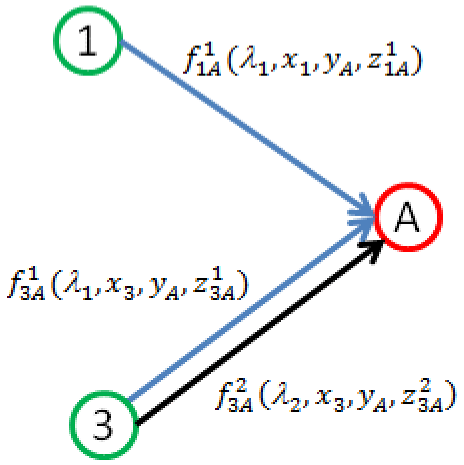

1
$f_{1A}^{1}(\lambda_1, x_1, y_A, z_{1A}^{1})$
A
$f_{3A}^{1}(\lambda_1, x_3, y_A, z_{3A}^{1})$
3
$f_{3A}^{2}(\lambda_2, x_3, y_A, z_{3A}^{2})$

FIGURE 5-5: Example of mode specific link based functions to predict the number of infections and vectors imported to susceptible node A, attributed to each adjacent endemic region (1 and 3)

The most critical issue is again determining the functional form of $f_{ji}^{k}(\lambda_k, x_j, y_i, z_{ji}^{k})$. Under the assumption that only infected vectors are transported via maritime cargo routes, and infected humans via air passenger routes, the mode specific sub-networks can be calibrated separately in the same manner as the single mode model. For the passenger travel network, the sub-model will be validated using regional reported traveler infection data. For the cargo network, the sub-model will be validated using reported vector population data.

Using the sample network in Figure 5-5, for cargo, defined as mode *k* = *1*, the total predicted number of infected vectors imported at *A* is $P_A^1 = \sum_{\forall j \in A_1(A)} f_{jA}^{1}(\lambda_1, x_j, y_A, z_{jA}^{1}) = f_{1A}^{1}(\lambda_1, x_1, y_A, z_{1A}^{1}) + f_{3A}^{1}(\lambda_1, x_3, y_A, z_{3A}^{1})$ where $A_k(A)$ represents the set of endemic nodes adjacent to *A* for mode *k*. For passenger air travel, defined as mode *k* = *2*, the predicted number of infected passengers imported at *i* is $P_A^2 = \sum_{\forall j \in A_2(A)} f_{jA}^{2}(\lambda_2, x_j, y_A, z_{jA}^{2}) \ = f_{3A}^{2}(\lambda_2, x_3, y_A, z_{3A}^{2})$.

The objective is then to provide an estimate of the overall risk that a specific source-destination pair presents as a combination of its passenger travel based risk and its cargo based risk. To accomplish this, mode-specific risk estimates for each link need to be aggregated over all available modes. Different approaches should be explored for this purpose, including weighted averages based on experimental data, (i.e. $F_{ji} = \sum_{\forall k} w_k f_{ji}^k$, where w_k is the weight associated with mode k, and F_{ji} is the aggregated risk posed by link (j,i)), and multi-criteria analysis methods.

If the individual modes cannot be calibrated independently (e.g. multiple modes transporting infected individuals) the original methodology will not apply, and a new functional form and calibration method needs to be defined. For either case the interaction between the two networks needs to be defined as well. This is an ongoing research effort.

5.8.3 Potential Extension to bi-level analysis

An additional level of analysis planned addressed real-time outbreak data (at an endemic source) and local climate conditions (at the susceptible destinations). This analysis evaluates route origins on an individual bases (selected based on the existence of an outbreak), and re-evaluates the relative risk of outgoing air travel routes using both the previously calibrated risk in conjunction with real-time local climate conditions at the destinations. The climate conditions are assessed using remote satellite imagery which captures features such as the presence of standing water, the most significant factor for mosquito breeding. This second level of analysis supports proactive mitigation strategies, (i.e. mosquito control efforts can be better informed).

CHAPTER 6: CONCLUSIONS AND FUTURE RESEARCH

Modeling the spatiotemporal spread of infectious disease is a multi-faceted problem. The stochastic nature of infection dispersal and the interdisciplinary nature of the problem present additional challenges to accurately depicting the epidemiological process. These issues are addressed in this dissertation in three different network based models for predicting infection spreading patterns. The network structures were derived from human mobility patterns, with a strong emphasis on passenger air travel. The methodologies have the potential to be extended to other transportation based systems independently, as well as multimodal systems. Both human contact-based and vector-borne diseases were addressed. The main contribution is the incorporation of dynamic infection data, which is becoming increasingly available; differentiating the proposed models from probabilistic epidemiological models which predict *expected* outbreak patterns and properties for a potential future outbreak. The models presented in this dissertation exploit infection data among other network properties to identify the spatiotemporal outbreak pattern for a *specific* spreading scenario.

6.1 Overview of Dissertation

The first chapter motivated the development of models for predicting the role of transportation in the spread of infectious diseases. The broad range of potential applications for such a model was revealed through a selection of network-based processes which exhibit similar behavioral characteristics to infection dispersal. Chapter 2 reviewed basic network structures and corresponding properties, followed by the most current research methods for predicting disease spreading in networks. Both microscopic (regional) and macroscopic (inter-regional) level models were reviewed, as well as some integration techniques for the two.

Chapters 3-5 addressed three different applications of infection dispersal on networks, increasing in scope from the community level modeled using social contact networks, to the international level modeled using passenger air traffic networks. Each chapter introduced a different problem, and presented a mathematical definition, solution methodology, problem application and numerical results. Each solution methodology exploits the use of spatiotemporal infection data. The problems introduced in chapter 3 and 4 invoke a similar solution methodology to infer the most likely spatiotemporal path of infection spanning from a single infection source, but vary in their applications. In chapter 3 the methodology was implemented on a social network defined by individual contact patterns. In chapter 4 the problem scope was increased; the new network is derived from inter-regional human travel patterns, specifically passenger air travel data. The problem introduced in chapter 5 varies significantly from the prior two, and introduced a new solution methodology, new network structure, and inherently different infection spreading process (which involves a third climate sensitive spreading agent). The solution methodology identifies the highest risk international air travel routes in terms of importing vector-borne diseases.

Each of the models contributes to the development of real-time analysis and decision support for ongoing outbreak scenarios.

6.2 CONTRIBUTIONS

The models introduced in this dissertation contribute towards various inter-related fields of study. Transportation modeling represents the core of this research. The network structures analyzed specifically exploit human mobility patterns, and are derived from various transportation systems. The main emphasis is on air-travel networks, but alternative modes of transportation can also be analyzed using similar methodology. The role of transportation in spreading infectious disease is of increasing importance as regional and global transportation systems continue to expand geographically, and increase in speed, efficiency and use.

A contribution specific to the epidemiological literature is the incorporation of real-time information into network based prediction models. A more abstract contribution of this research falls under the umbrella of complex systems. Representing the disease dispersal process requires modeling the interaction among multiple network based systems (e.g. social-contact networks, transportation-based network systems, ecological-geographic spatial networks). Some simplified examples of multi-layered network applications are introduced throughout the dissertation, such as the multimodal extensions in chapter 4 and 5. This is a research topic which will continue to be expanded on in the future.

6.3 CRITICISMS

Incomplete infection data is currently the most limiting factor in the potential performance of the models. For example, in chapter 4 and 5 the limited data necessitated a level of aggregation resulting in unrealistic assumptions and problem properties. A lack of data also makes it is difficult to truly asses the predictive capability of the models. One major motivation for the development of these models is to incentivize better data collection efforts.

For the models introduced in this dissertation route-level infection data would be the most valuable to collect. This type of information requires information on infected individuals and their recent travel history. Route level data would allow quantitative analysis of the models' performance. In addition to route level data, more (spatially and temporally) disaggregated infection data is necessary to implement various proposed extensions to the models.

6.4 FUTURE RESEARCH DIRECTION:

Enhanced data is one possibility for improving model performance. Additionally, each model has the potential to be expanded methodologically in multiple directions. Extensions specific to each model are introduced in their respective chapter, however a reoccurring theme across chapters addresses some form of interdependent network analysis. This fundamental research topic is expanded upon in the following section.

6.4.1 Interdependent Network Analysis

To realistically model processes which bridge multiple network systems it is necessary to define system interdependencies. Related examples of interdependent systems (within the realm of disease spreading prediction models) include:

i. Multi-modal transportation systems (air, cargo, shipping, freight, etc.)
ii. Human mobility (Transport) networks spanning a geographic spatial grid

Network models representative of these unified systems need to be developed and analyzed with respect to their impact on the infection spreading process. To accomplish this, characteristics of a *global* network structure need to be identified, where *global* refers to the multi-dimensional (coupled) system. The fundamental questions which must be answered include:

i. How do we define a coupled system?
ii. How do we represent it?
iii. How do we construct it?
iv. How does the coupled system behave in comparison to the systems independently in terms of structural properties:
    a. Size, degree distribution, connectivity, etc?

Answers to these questions will inevitable vary dependent on the application. A simple example of coupling two different networks is motivated from Problem III. As it is currently defined, Problem III represents a highly simplified coupling of the air transportation network and a geographic spatial network, where all the attributes of the spatial network have been condensed to a single representative value (e.g. suitability). Disaggregating the spatial network such that it has an explicitly defined network structure with link and node properties adds a new dimension to this analysis.

Two network systems which concurrently contribute to the dispersal of a biological infecting agent, such as a mosquito, are a geographic spatial network and a transport network (i.e. air traffic network, maritime cargo network). For example, a mosquito has a limited ability to travel independently (fly) across geographic space, but can also be transported across larger distances via transport links (such as a plane, cargo ship, etc.). Modeling the interaction of these two networks is integral is mapping the potential spread of a given biological infecting agent. In the example in this section the geographic network structure is represented as a square grid, defining a set of discrete spatial regions. The transport system is defined using a heterogeneous network structure. Figure 6-1 illustrates examples of these two network structures, including how the set of geographic links is defined. By representing each geographic region as a node, linking nodes in adjacent regions form potential (flight) paths through geographic space. Diagonal links may also exist, but they are not included in this example for simplicity.

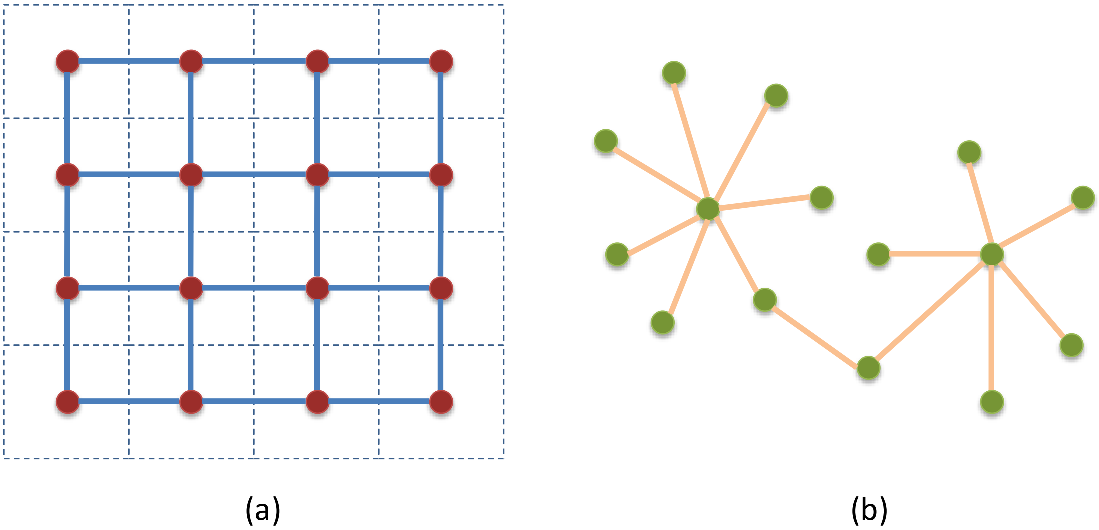


FIGURE 6-1: Example of a (a) geographic spatial grid and (b) transport network structure

There are various alternatives for coupling two networks, which will vary based on the application and network properties. The most basic question is how to integrate the two layers? This can be accomplished by defining new connections between pairs of nodes. The following steps illustrate a possible method for defining these connections, and creating a global network structure.

I. Assign each node in the transport network to a single geographic region (this could be based on an airport location). The regional assignment is random in this example, and illustrated in figure 6-2.

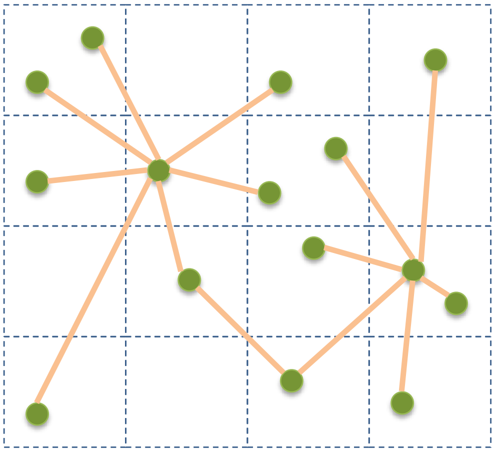


FIGURE 6-2: Spatial assignment of transport network

II. Add new spatial connections (blue) between transport nodes based on geographic proximity, see figure 6-3.

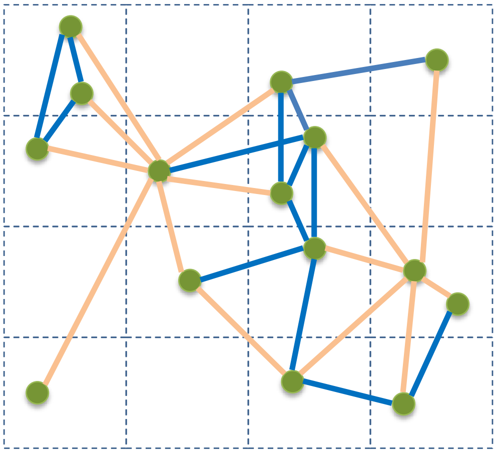


FIGURE 6-3: Additional direct transport network connections due to spatial proximity

III. Add new transport connections (orange) between geographic regions based on the transport network structure, see figure 6-4.

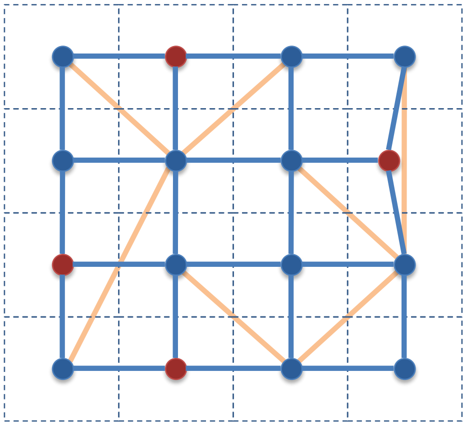


FIGURE 6-4: Additional direct geographic spatial connections due to transport links

Figures 6-5 represents how the two example network structures might overlap in space, detailing the new inter-regional spatial connections between different node types.

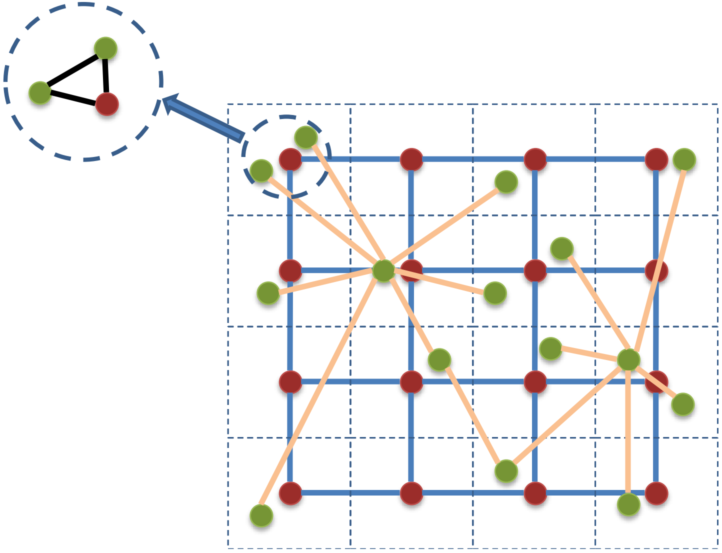


FIGURE 6-5: Spatial overlap of networks

In figure 6-3 the new spatial connections in the transport network are defined by connecting any pair of transport nodes located in either the same or adjacent geographic region. This type of link might represent a mosquito's potential flight path (e.g. between airports). For this type of application the connection rules should be a function of both the maximum distance a mosquito can fly and the scale of the geographic regions. The blue links in figure 6-3 represent these new connections between transport nodes due to their spatial proximity. The original transport connections remain constant.

Similarly, the introduction of the transport network may increase the connectivity of the geographic grid by introducing routes capable of carrying a mosquito across larger geographic spaces. In this example the new transport links are defined by connecting any two geographic regions *a* and *b* for which a transport node in *a* is directly connected to a transport node in *b*. This type of link may represent the possible transmittal of an infected mosquito into a new region via an airplane (or cargo ship). Again the inclusion of these links might be based on the maximum length flight (or cargo route) a mosquito is capable of surviving. The increased connectivity is illustrated in figure 6-4, where the orange links are new transport connections between geographic nodes. The original geographic connections remain constant. The blue nodes are the regions where transport nodes were assigned.

The previous example represented one method for integrating two network structures. An alternative simplified method is implemented to illustrate how network properties (degree distribution, shortest path length) are affected by the coupling process. For this example the same two networks shown in figure 6-1 are coupled, but the only added links are between a transport node and the geographic node representing the region to which it is assigned. All other links remain constant. The resulting global network structure is illustrated in figure 6-6, where the black links are the newly added connections. This method results in fewer new connections than the first method presented, but still results in a significant impact on the network properties. (This might be a more appropriate integration method when the geographic regions are larger.)

For both methods the global network structure introduces a new set of paths connecting node pairs. New links connect geographic nodes directly to transport nodes, providing a means to model interaction between these previously independent systems. Additionally, both initial independent networks may increase in connectivity.

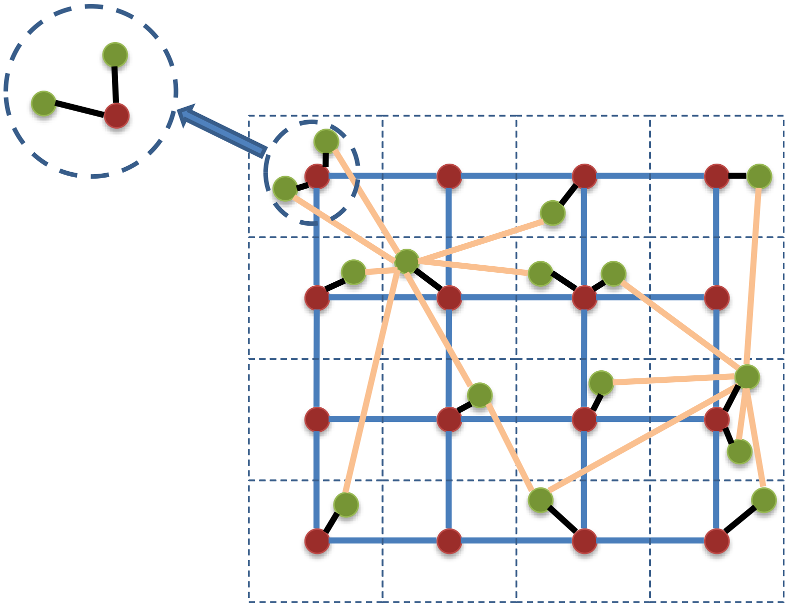


FIGURE 6-6: Global network: Transport nodes are connected to assigned regional node

Introducing highly connected transport nodes into a geographic region may increase the connectivity of the region. A simple illustration of this is shown in figure 6-

7. The transport nodes assigned to regions retain their original connections, thus increasing the region degree. For this example the transport nodes are randomly assigned to regions, so the number of transport nodes in each region is uniformly distributed, but the degree distribution remains heterogeneous. Therefore the initially uniform regional degree distribution is significantly increased for certain regions. Figure 6-8 illustrates this change in regional degree distribution when the transport network structure is power law, and has 600 nodes. The impact on regional degree depends on the connectivity rules.

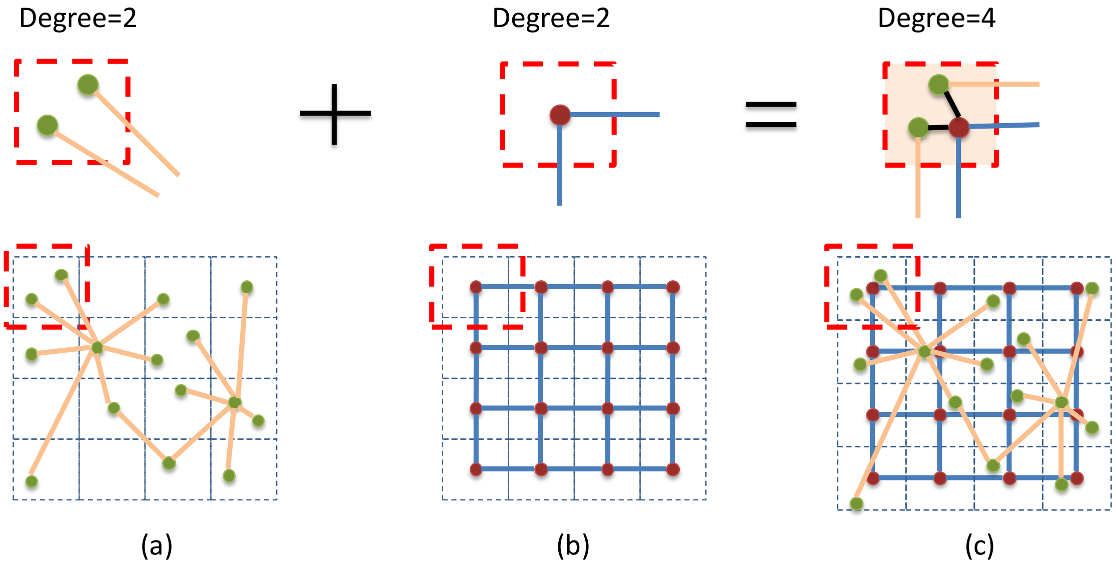


FIGURE 6-7: Example calculation of (c) global region degree after coupling (a) transport network and (b) geographic grid

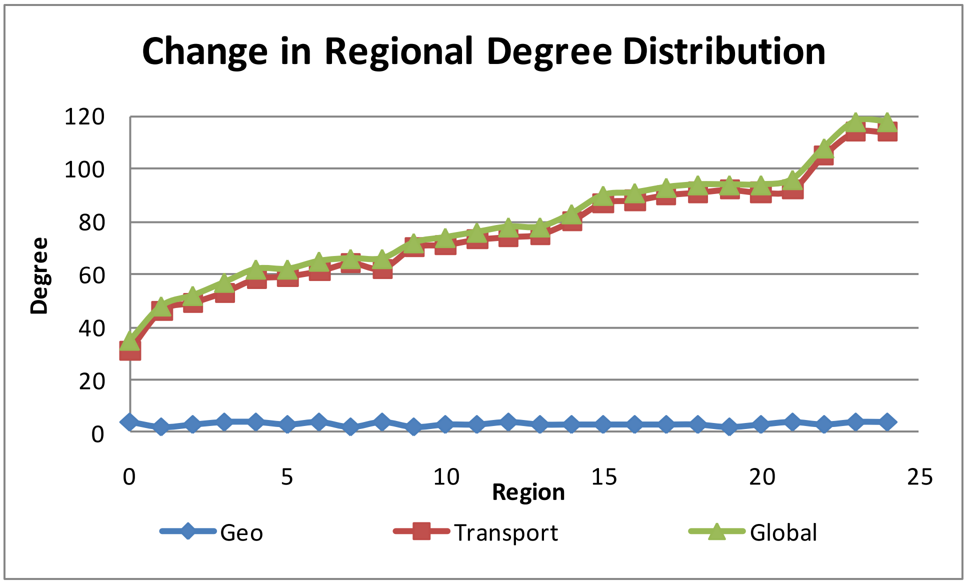


FIGURE 6-8: For 5x5 Geographic Grid, example of increased regional degree due to coupling with transport network

In addition to creating paths between previously disconnected node sets, coupling two networks increases the connectivity of each system. Because all three network structures (transport, geographic, global) are connected (a path exists between all node pairs) it is possible to calculate a shortest path cost between all pairs of nodes. The average shortest path cost (averaged over all node pairs) is one parameter which characterizes the connectivity of a network. In this analysis the increase in connectivity for a network is formally defined as the percent decrease in the average shortest path cost between all pairs of nodes (in a given node set) before and after coupling the systems. The two figures below illustrate this effect for both the geographic and transport networks. For the transport network, this is the decrease in average shortest path cost between all transport nodes (although geographic nodes can be included in the paths) using the global set of links, relative to the original set of transport links. The geographic case is calculated analogously.

The figures 6-9 and 6-10 illustrate the increased connectivity of each network system as a function of transport network size (number of nodes), for both a 5x5 geographic grid (figure 6-9) and a 10x10 geographic grid (figure 6-10). The results indicate transport connections have more impact on a larger geographic grid, intuitively providing more significant shortcuts between regions. As the transport network increases in size, the geographic connectivity increases, but stabilizes after the transport network reaches a certain size. Similarly, the transport system becomes more connected with the availability of the geographic links. For the smaller geographic grid the transport network benefits more from the geographic "short cuts" than the geographic network benefits from transport connections. This is likely a function of the ratio of the size of the transport network compared to the geographic space; the same result does not occur on the 10x10 grid.

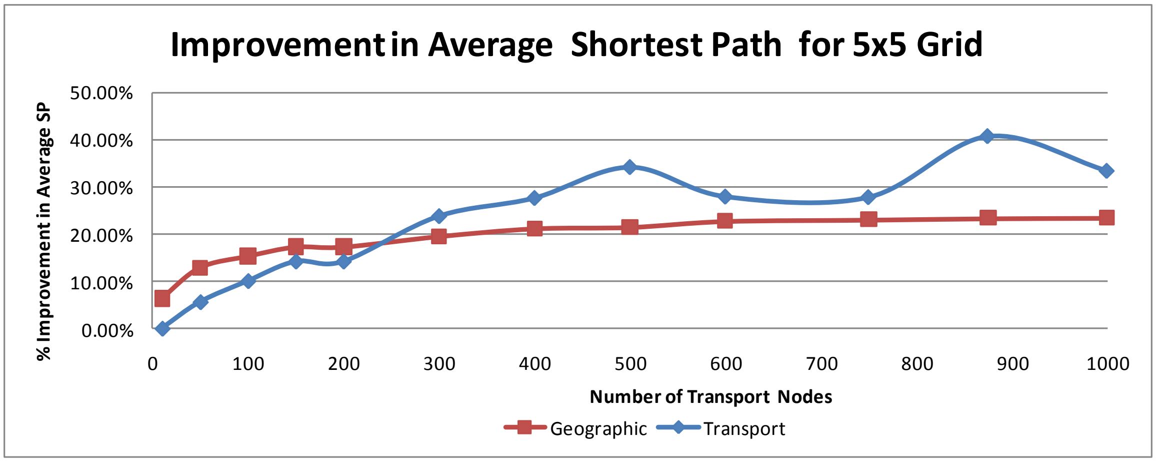


FIGURE 6-9: Percent decrease in average shortest path as size of transport network increases for 5x5 grid

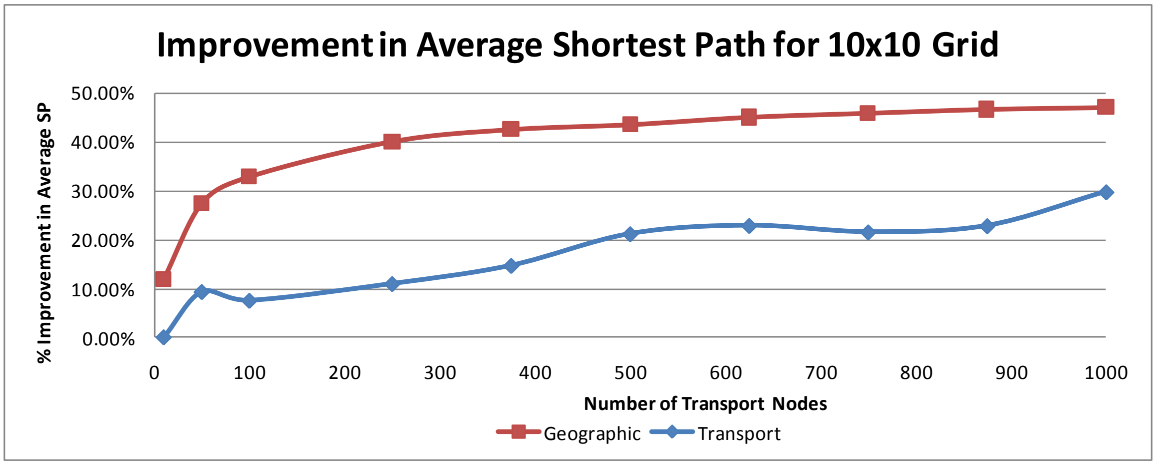


FIGURE 6-10: Percent decrease in average shortest path as size of transport network increases for 10x10 grid

In this example the links all have a unit cost, so the shortest path is equivalent to the minimum hop path. In addition, the link types are indistinguishable. An extension of this analysis should differentiate between link types and define link costs as a function of link and system properties. Analysis for alternative network structures should also be explored.

Coupling network systems is essential to modeling the interaction and dependencies amongst those systems. In this dissertation the focus is the combined impact of various network systems on escalating the spread of infectious disease. An analysis accounting for network interdependencies in the spread of infectious disease may expose new effective mitigation strategies. In order to reveal the complex dispersal characteristics of infection future epidemiological models should incorporate such system interdependencies.

REFERENCES

1) "1918 Influenza Pandemic | CDC EID". Archived from the original on 2009-10-01. Retrieved 2009-09-28.

2) Agnati, L. F., Guidolin, D. and Fuxe, K.. (2007). "The brain as a system of nested but partially overlapping networks. Heuristic relevance of the model for brain physiology and pathology"*, Journal of Neural Transmission* 114 (1): 3-19.

3) Albert, R. and Barabási, A.-L. (2002). "Statistical mechanics of complex networks", Reviews of Modern Physics 74: 47–97.

4) Atkinson, M.P. and Wein, L.M. (2009). "An Overlapping Networks Approach to Resource Allocation for Domestic Counterterrorism", *Studies in Conflict & Terrorism*, 33(7): 618 – 651.

5) Bagrow, JP and Bollt, EM. (2005)**. "**Local method for detecting communities", *Phys. Rev. E 72*.

6) Balcan D, Colizza V, Goncalves B, Hu H, Ramasco JJ, et al. (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. *P Natl Acad Sci* USA 106: 21484–21489.

7) Barabási, Albert-László and Albert, Réka. (1999). "Emergence of scaling in random networks". *Science* 286: 509–512.

8) Bar-Yam, Y. (1997). Dynamics of complex systems. Chapter 2, section 3. Reading, Mass., Addison-Wesley.

9) Benedict, Brandy (2007). "Modeling Alcoholism as a Contagious Disease: How "Infected" Drinking Buddies Spread Problem Drinking", *SIAM News*, 40 (3).

10) Bianconi, G., Gulbahce, N. and Motter**,** A. E. (2008). "Local structure of directed networks", *Phys. Rev. Letters* 100, 118701.

11) Brockmann, D. (2008). "Theis: Money Circulation, Trackable Items, and the Emergence of Universal Human Mobility Patterns**"**. *Pervasive Computing* **7** (4).

12) Brockmann, D. Hufnagel, L. and Geisel, T. (2006). "The scaling laws of human travel" *Nature* 439, 462.

13) Brockmann, D. (2008). "Anomalous diffusion and the structure of human transportation networks", *European Physical Journal - Special Topics* **157**, 173-189.

14) Brockmann, D.. (2009). "Human Mobility and Spatial Disease Dynamics", *Reviews of Nonlinear Dynamics and Complexity*, H. G. Schuster (ed.), Wiley-VCH.

15) Brownstein JS, Wolfe CJ, Mandl KD. (2006). "Empirical Evidence for the Effect of Airline Travel on Inter-Regional Influenza Spread in the United States". *PLoS Med* 3(10): e401.

16) CDC. (2005). "Travel-Associated Dengue Infections --- United States, 2001—2004", *MMRW,* 54(22), 556-558.

17) CDC. (2007). "Dengue hemorrhagic fever—U.S.-Mexico border, 2005". *MMWR*, 56(31), 785-9.

18) CDC. (2010). "Locally Acquired Dengue --- Key West, Florida, 2009—2010", *MMRW*, 59(19), 577-581.

19) Christakis, N.A. and Fowler, J.H. (2007). "The Spread of Obesity in a Large Social Network over 32 Years", *N Engl J Med*; 357:370-379

20) Colizza V, Barrat A, Barthélemy M, Vespignani A. (2006). "The modeling of global epidemics: Stochastic dynamics and predictability", *Bull Math Biol* 68: 1893–1921.

21) Colizza V, Barrat A, Barthélemy M, Vespignani A. (2006). "The role of the airline transportation network in the prediction and predictability of global epidemics", *Proc Natl Acad Sci U S A* 103: 2015–2020.

22) Cooper BS, Pitman RJ, Edmunds WJ, Gay NJ. (2006). "Delaying the International Spread of Pandemic Influenza", *PLoS Med* 3(6): e212.

23) Cooper BS, Pitman RJ, Edmunds WJ, Gay NJ. (2006). "Delaying the International Spread of Pandemic Influenza". *PLoS Med* 3(6): e212.

24) Cottam *et al.*, (2008). "Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus", *Proc. R. Soc. B,* 275, 887-895

25) Crucitti, P., Latora, V. and Marchiori, M. (2004). "Model for cascading failures in complex networks", *Phys. Rev. E*, 69(4):045104.

26) Disease Vectors Database. www.diseasevectors.org. [Last Accessed 28 February 2010].

27) Drummond AJ, Rambaut A. (2007). "Beast: Bayesian evolutionary analysis by sampling trees", *BMC Evol Biol* 7: 214.

28) Dueñas-Osorio, Leonardo, Craig, J.I. and Goodno, B.G. (2004). "Probabilistic response of interdependent infrastructure networks", *Proceedings of the 2nd annual meeting of the Asian-pacific network of centers for earthquake engineering research (ANCER)*. Honolulu, Hawaii. July 28-30.

29) Dueñas-Osorio. (2005). “Interdependent Response of Networked Systems to Natural Hazards and Intentional Disruptions”. PhD Dissertation submitted to Georgia Institute of Technology.

30) Edmonds J. (1967). “Optimum Branchings”. *J. Res. Nat. Bur. Standards* 9, 233-240.

31) Elith J, Graham CH, Anderson RP, Dud´ık M, Ferrier S, et al. (2006). Novel methods improve prediction of species’ distributions from occurrence data. *Ecography*, 29, 129-151.

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

33) Erdős, P.; Rényi, A. (1960). "The Evolution of Random Graphs". *Magyar Tud. Akad. Mat. Kutató Int. Közl.* 5: 17–61.

34) Eubank, S. (2005). “Network Based Models of Infection Disease Spread”, *Jpn. J. Infect. Dis*., 58.

35) Eubank, S., Guclu, H., et al. (2004). “Modeling disease outbreaks in realistic urban social networks”. *Nature* 429, 180–184.

36) European Center for Disease Control (ECDC). http://www.ecdc.europa.eu/ [Accessed 2010].

37) Eurostat: European Commission Statistics Database. http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home [Last Accessed July, 2010].

38) Fainzang, Sylvie. (1996). “Alcoholism, a contagious disease. A contribution towards an anthropological definition of contagion”, *Culture, Medicine and Psychiatry*, 20 (4), 473-487.

39) Federal Aviation Administration: Passenger Boarding (Enplanement) and All-Cargo Data for U.S. Airports. http://www.faa.gov/airports/planning_capacity/passenger_allcargo_ stats/passenger/index [Last Accessed July, 2010]

40) Franklin, J. (2009). Mapping Species Distributions: Spatial Inference and Prediction. *Cambridge University Press*.

41) Freedman, DO, *et al*. (2006). "Spectrum of disease and relation to place of exposure among ill returned travelers". *The New England Journal of Medicine*, 354(2), 119-130.

42) Girvan, M and Newman, MEJ. (2002) "Community structure in social and biological networks" *PNAS*.

43) González C, Wang O, Strutz SE, González-Salazar C, Sánchez-Cordero V, et al. (2010). Climate change and risk of Leishmaniasis in North America: Predictions from ecological niche models of vector and reservoir species. *PLoS Negl Trop Dis,* 4: e585.

44) Grais RF, Ellis JH, Kress A, Glass GE. (2004). "Modeling the spread of annual influenza epidemics in the U.S.: The potential role of air travel", *Health Care Manag Sci* 7: 127–134.

45) Grassberger P. "On the critical behavior of the general epidemic process and dynamical percolation", *Math.Biosci*. 63: 157-162, 1983.

46) Guare, John (1990). *Six Degrees of Separation: A Play* (First edition ed.). New York: Random House

47) Gubler D, Kuno G. (1997). "Dengue and dengue hemorrhagic fever: its history and resurgence as a global public health problem", *London: CAB International*, 1–22.

48) Gubler, D.J., Reiter, P., Ebi, K.L., Yap, W., Nasci, R., Patz, J.A. (2001). “Climate variability and change in the United States: Potential impacts on vector-and rodent-borne diseases”, *Environmental Health Perspectives*, 109(2), 223-233.

49) Halloran, M. Elizabeth *et al.* (2008). *“*Modeling targeted layered containment of an influenza pandemic in the United States”, *PNAS* 105 (12): 4639–4644.

50) Haydon D.T, Chase-Topping M, Shaw D.J, Matthews L, Friar J.K, Wilesmith J, Woolhouse M.E.J (*2003). “*The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak*”, Proc. R. Soc. B. 270, 121–127.*

51) Hays, J. N. (2005). "Epidemics and pandemics: their impacts on human history". *ABC-CLIO, Inc*. p.23.

52) Hijmans R, Cameron S, Parra J, Jones P, Jarvis A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25, 1965-1978.

53) Hufnagel, L., Brockmann, D. and Geisel, T.. (2004). “Forecast and control of epidemics in a globalized world”, *Proc Natl Acad Sci USA* 101, 15124.

54) Jelinek T, *et al*. (2002). “Epidemiology and clinical features of imported dengue fever in Europe: sentinel surveillance data from TropNetEurop”, *Clin Infect Dis*., 35(9), 1047-52.

55) Jelinek, T. (2009). “Trends in the epidemiology of dengue fever and their relevance for importation to Europe”, *Eurosurveillance*, 14 (25): Article 2.

56) Jombart, Thibaut; Eggo, Rosalind M; Dodd, Pete; Balloux, Francois. (2009). “Spatiotemporal dynamics in the early stages of the 2009 A/H1N1 influenza pandemic”. *PLoS Currents Influenza*. RRN1026.

57) Kaluza, Pablo, Kölzsch, Andrea, Gastner, Michael T. and Blasius, Bernd. (2010). The complex network of global cargo ship movements. *J. R. Soc. Interface 7:1093-1103; doi:10.1098/rsif.2009.0495*

58) Keeling, M, and Eamesm K.T. (2005). "Networks and Epidemic Models", *J. R. Soc. Interface* 2 (4): 295-307.

59) Knowles L, Maddison W. (2002). Statistical phylogeography. *Molecular Ecology* 11: 2623–2635.

60) Lemey P, Rambaut A, Drummond AJ, Suchard MA. (2009). "Bayesian Phylogeography Finds Its Roots", *PLoS Comput Biol* 5(9).

61) Lemey P, Suchard M, Rambaut A. (2009). "Reconstructing the initial global spread of a human influenza pandemic: A Bayesian spatial-temporal model for the global spread of H1N1pdm". *PLoS Curr Influenza*.

62) Longini, IM Jr, Fine PE, Thacker SB. (1986). "Predicting the global spread of new infectious agents", *Am J Epidemiol* 123: 383–391.

63) Mairuhu, A. T., J. Wagenaar, D. P. Brandjes, and E. C. van Gorp. (2004). "Dengue: an arthropod-borne disease of global importance", *Eur. J. Clin. Microbiol. Infect. Dis.* 23, 425-433.

64) Margules CR, Sarkar S (2007). Systematic Conservation Planning. Cambridge, UK: *Cambridge University Press.*

65) Massad, Eduardo and Wilder-Smith, Annelies. (2009). "Risk Estimates of Dengue in Travelers to Dengue Endemic Areas Using Mathematical Models", *Journal of Travel Medicine*, 16 (3): 191–193.

66) *MC Gonzalez, CA Hidalgo, A-L Barabasi* (2008). "Understanding Individual Human Mobility Patterns" *Nature* 453: 779-782

67) Meyers L., Pourbohloul, B., Newman, M. E. J., Skowronski, D. and R. Brunham. "Network theory and SARS: Predicting outbreak diversity", *Journal of Theoretical Biology* 232, 71–81, 2005.

68) Milgram, Stanley. (1967). "The Small World Problem", *Psychology Today,* 1(1): 60–67.

69) Milo, R. et al. (2002)**. "**Network Motifs: Simple Building Blocks of Complex Networks", *Science* 298:824-827

70) Morens, DM, Fauci AS. (2008). "Dengue and Hemorrhagic Fever: A Potential Threat to Public Health in the United States", *JAMA*, 299(2), 214-216.

71) Moffett A, Shackelford N, Sarkar S (2007) Malaria in Africa: Vector species niche models and relative risk maps. *PLoS ONE* 2: e824.

72) Moffett A, Strutz S, Guda N, Gonz´alez C, Ferro MC, et al. (2009). A global public database of disease vector and reservoir distributions. *PLoS Negl. Trop. Dis*. 3: e378.

73) Newman M. E. J. "Models of the small world", *J. Stat. Phys.* 101, 819–841, 2000.

74) Newman M. E. J., Strogatz, and D. J. Watts. (2001). "Random graphs with arbitrary degree distributions and their applications", *Physical Review. E* 64, 026118.

75) Newman M.E.J. (2002). "Spread of epidemic disease on networks", *Physical Review E* 66, 016128.

76) Newman, M. E. J., Forrest, Stephanie and Balthrop, Justin (2002). "Email networks and the spread of computer viruses", *Physical Review E* 66.

77) Newman, M.E.J. and Watts , D. J. (1999). "Scaling and percolation in the small-world network model", *Physical Review E* 60, 7332–7342.

78) Palla, G., Barabasi, A. L. & Vicsek, T. (2007). "Quantifying social group evolution", *Nature,* 446, 664-667.

79) Palla, G., Derenyi, I., Farkas, I. & Vicsek, T. (2005). "Uncovering the overlapping community structure of complex networks in nature and society", Nature, 435, 814-818.

80) Peterson AT (2008) Biogeography of disease: a framework for analysis. Naturwissenschaften 95: 483–491.

81) Phillips SJ, Dudìk M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, 31, 161-175.

82) Ravasz, E. *et al.* (2002). **"**Hierarchical organization of modularity in metabolic networks", *Science* 297, 1551-1555.

83) Rigau-Perez, J. G., Gubler, D. J., Vorndam, A. V., and Clark, G. (1997). Dengue: a literature review and case study of travelers from the United States, 1986-1994. *Journal of Traveling Medicine,* 4, 65-71.

84) Ronald Ross. (1916). "An Application of the Theory of Probabilities to the Study of a priori Pathometry. Part I", *Proc. R. Soc. Lond. A,* 92:204-230.

85) Rvachev L, Longini I. (1985). "A mathematical model for the global spread of influenza", *Math Biosci* 75: 3–22.

86) Sachs, Jeffrey D. (2008). "Blackouts and Cascading Failures of the Global Markets: Feedbacks in the economic network can turn local crises into global ones", *Scientific American Magazine.*

87) Saramäki, J. *et al.* (2007). “Generalizations of the clustering coefficient to weighted complex networks”, *Phys. Rev. E* 75, 027105

88) Sarkar, S., Strutz, S., Frank, D. M., Rivaldi, C. –L., and Sissel, B. 2010. “Chagas Disease Risk in Texas.” *PLoS Neglected Tropical Diseases* 4 (10): e836. doi:10.1371/journal.pntd.0000836.

89) Small M. and C.K. Tse. (2005). “Clustering model for transmission of the SARS virus: application to epidemic control and risk assessment”, *Physica A* 351, 499–511.

90) Small M. and C.K. Tse. (2005). “Small world and scale free model of transmission of SARS”, *International Journal of Bifurcations and Chaos*.

91) Small M., Shi, P. and C.K. Tse. (2004). “Plausible models for propagation of the SARS virus”, *IEICE Trans. Fund. Electron. Commun. Comput. Sci.* E87-A 2379–2386.

92) Standish K, Kuan G, Avilés W, Balmaseda A, Harris E. (2010). “High Dengue Case Capture Rate in Four Years of a Cohort Study in Nicaragua Compared to National Surveillance Data”, *PLoS Negl Trop Dis*, 4(3), e633.

93) Suchard MA, Rambaut A. (2009). Many-core algorithms for statistical phylogenetics. Bioinformatics., 25(11): 1370-6.

94) Tatem AJ, Hay, SI, Rogers DJ (2006). “Global traffic and disease vector dispersal”, *PNAS,* 103(16), 6242–6247.

95) Tatem, A. J. and Hay, S. I. (2007). “Climatic similarity and biological exchange in the worldwide airline transportation network”. *Proc. R. Soc. B* (2007) 274: 1489–1496.

96) Tatem1, A.J., Rogers, D.J. and Hay, S.I. (2006). “Global Transport Networks and Infectious Disease Spread”, *Advances In Parasitology* 62.

97) Thadakamalla, H.P., Kumara, S. R. T. and Albert, R. (2007). “Complexity and Large-scale Networks”**,** Chapter 11 in *Operations Research and Management Science Handbook* edited by A. R. Ravindran, CRC press.

98) TropNetEurope: European Network on Imported Infectious Disease Surveillance. http://www.tropnet.net/index_2.html [Accessed on July 2010].

99) United States Census Bureau. (2010). "Resident Population Data: Population Change".http://2010.census.gov/2010census/data/apportionment-pop-text.php. Retrieved December 23, 2010.

100) U.S. Department of Transportation’s Research and Innovative Technology Administration (RITA). http://www.rita.dot.gov/about_rita/ [Last Accessed July, 2010].

101) US Department of Commerce, International Trade Administration. http://tinet.ita.doc.gov/view/m-2007-O-001/index.html [Accessed 2010].

102) Vazquez, A. et al. (2004). “The topological relationship between the large-scale attributes and local interactions patterns of complex networks”, *PNAS* 101:17940-17945.

103) Viboud C, Miller MA, Grenfell BT, Bjørnstad ON, Simonsen L. “Air Travel and the Spread of Influenza: Important Caveats”. PLoS Med 3(11): e503, 2006.

104) Viger, F. and Latapy, M. (2005). Efficient and simple generation of random simple connected graphs with prescribed degree sequence. In Proc. 11th Conf. on Computing and Combinatorics (COCOON), 440–449.

105) Wade, Nicholas (2010). "Europe’s Plagues Came From China, Study Finds". New York Times. http://www.nytimes.com/2010/11/01/health/01plague.html. Retrieved 2010-11-01.

106) Wallace, R.G., HoDac, H., Lathrop,R.H. and Fitch, W.M. (2007). “A statistical phylogeography of influenza A H5N1”, PNAS 104 (11): 4473–4478.

107) *Wang, P, Gonzalez, MC, Hidalgo, CA and Barabasi, A-L.* (2009). "Understanding the spreading patterns of mobile phone viruses", Science, 324:1071-1076.

108) Warren CP, Sander LM, Sokolov IM, Simon C, Koopman J. (2002). “Percolation on heterogeneous networks as a model for epidemics”, Math Biosci. 180:293–305.

109) Warren, D. L. and Seifert, S. N. (Under review). Environmental Niche Modeling using Maxent: Do Under- and Over-parameterization Matter, and What Can We Do about It? Ecography.

110) Watts, D.J.; Strogatz, S.H. (1998). "Collective dynamics of 'small-world' networks.". Nature 393 (6684): 409–10.

111) Wichmann, O., Mühlberger, N. and T Jelinek. (2003). Dengue – The Underestimated Risk in Travelers. Dengue Bulletin, 27, 126-137.

112) Wilder-Smith A, Schwartz E. (2005). “Dengue in Travelers”, NEJM, 353(9), 924-932.

113) Wilder-Smith, A., Gubler, D.J. (2008). “Geographic expansion of dengue: The impact of international travel”, Medical Clinics of North America, 92, 1377–1390.

114) WorldClim database. www.wordclim.org. [Last Accessed 28 February 2010].

115) World Health Organization: Severe Acute Respiratory Syndrome (SARS) [Last Accessed November, 2010].

116) World Health Organization: Dengue Bulletin [Last Accessed July, 2010].

117) World Health Organization – DengueNET. http://apps.who.int/globalatlas/default.asp [Last Accessed July, 2010].