Modeling Social Response to the Spread of an Infectious Disease

by

Jane A. Evans

B.S. Operations Research, U.S. Air Force Academy, 2010

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Masters of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Author .....
Sloan School of Management
May 18, 2012

Certified by .....
Dr. Natasha Markuzon
The Charles Stark Draper Laboratory, Inc.
Technical Supervisor

Certified by .....
Marta Gonzalez
Assistant Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by .....
Patrick Jaillet
Dugald C. Jackson Professor, EECS
Co-Director, Operations Research Center

Modeling Social Response to the Spread of an Infectious Disease

by

Jane A. Evans

Submitted to the Sloan School of Management
on May 18, 2012, in partial fulfillment of the
requirements for the degree of
Masters of Science in Operations Research

Abstract

With the globalization of culture and economic trade, it is increasingly important not only to detect outbreaks of infectious disease early, but also to anticipate the social response to the disease. In this thesis, we use social network analysis and data mining methods to model negative social response (NSR), where a society demonstrates strain associated with a disease. Specifically, we apply real world biosurveillance data on over 11,000 initial events to: 1) describe how negative social response spreads within an outbreak, and 2) analytically predict negative social response to an outbreak. In the first approach, we developed a meta-model that describes the interrelated spread of disease and NSR over a network. This model is based on both a susceptible-infective-recovered (SIR) epidemiology model and a social influence model. It accurately captured the collective behavior of a complex epidemic, providing insights on the volatility of social response. In the second approach, we introduced a multi-step joint methodology to improve the detection and prediction of rare NSR events. The methodology significantly reduced the incidence of false positives over a more conventional supervised learning model. We found that social response to the spread of an infectious disease is predictable, despite the seemingly random occurrence of these events. Together, both approaches offer a framework for expanding a society's critical biosurveillance capability.

Technical Supervisor: Dr. Natasha Markuzon
Title: Principal Member of the Technical Staff
The Charles Stark Draper Laboratory, Inc.

Thesis Supervisor: Marta Gonzalez
Title: Assistant Professor of Civil and Environmental Engineering
Massachusetts Institute of Technology

Acknowledgments

There are many people I wish to thank for helping me achieve this dream.

I am deeply grateful to my advisor Dr. Natasha Markuzon of Draper Laboratory. She has worked tirelessly to develop my abilities as a researcher and inspire me to achieve great things in life. I would not have been able to complete this thesis without her guidance, and will forever remember to "tell a story".

Additionally, I would like to thank Professor Marta Gonzalez for accepting a student so late in the research process. Her steadfast support made this thesis a reality.

Thank you to my friends and study group here at MIT. The friendship and academic support of Brian, Matt, Mike, and Andy has meant the world to me.

Finally, I need to thank my family, for always pushing me to excel in everything I do. Thank you to my mother for also being my friend when I needed one, and to my father, MLFY.

Publication of this thesis does not constitute approval by Draper or the sponsoring agency of the findings or conclusions contained herein. It is published for the exchange and stimulation of ideas.

The views expressed in this article are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, we describe a technical approach for modeling social response to the spread of an infectious disease. This chapter provides some background on the field of biosurveillance, lists the research objectives, outlines the technical approaches utilized, and presents a brief overview of the thesis organization.

1.1 Biosurveillance

Biosurveillance is a process for identifying disease in people, plants and animals. More specifically, it monitors global health and disease trends to identify emerging problems. This paper focuses on the operational side of biosurveillance, using real-time social reporting to anticipate, detect, recognize and track infectious disease events, and to facilitate a proactive response. However, there are a variety of other relevant biosurveillance definitions.

Biosurveillance also keeps track of viruses, bacteria and other agents that can cause disease, and identifies problems with public health infrastructure that can worsen a disease outbreak. A good biosurveillance system detects factors that predispose to disease, identifies cases of disease, predicts whether an outbreak or epidemic will occur, and anticipates secondary problems associated with an outbreak. Biosurveillance is a repetitive process of collecting and analyzing data, making decisions, and responding appropriately to a bio-event, as shown in Figure 1-1 [1].

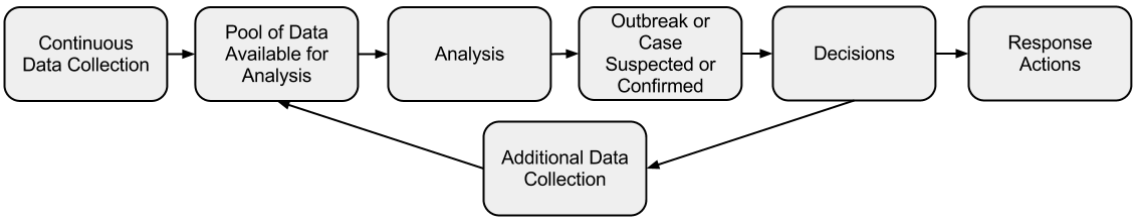


Figure 1-1: The biosurveillance process, represented as a positive feedback loop

The Centers for Disease Control and Prevention (CDC) has a more explicit definition of biosurveillance that emphasizes the response to an outbreak. The underlying purpose of biosurveillance is to prevent or alleviate the impact of an infectious disease outbreak on the population, such as malady, loss of life and economic impacts. Accordingly, the CDC defines biosurveillance as the practice of managing health-related data and information, to provide [2]:

- Early warning of threats and hazards
- Early detection of events
- Quick characterization of events to mitigate adverse health effects

Biosurveillance not only monitors natural disease outbreaks, but also incidents of bioterrorism. The CDC defines bioterrorism as the deliberate release of viruses, bacteria, or other agents used to cause illness or death in people, animals, or plants [3]. Because biological agents are extremely difficult to detect and have a delayed impact on the population, they are an ideal weapon for terrorists. The most effective way to detect a bioterrorism attack is through biosurveillance, which should identify an outbreak regardless of its origin. Thus biosurveillance is an crucial component of any modern government's defense program.

However, it is only recently that biosurveillance has become a viable option. As with many other types of surveillance, it depends on the collection and analysis of vast amounts of data. With the exponential rise in computing power, it is finally possible to identify emerging problems in a timely manner. Yet despite the increase in biosurveillance data, a unified approach is still lacking. As an example, consider the

June 2011 outbreak of enterohaemorrhagic *E. coli* in Germany. The outbreak spread to most of Europe and underscored the need for increased biosurveillance capability. Considered one of the world's worst outbreaks of *E. coli*, the deadly strain infected over 3000 people and caused at least 39 deaths [4]. Figure 1-2 shows the extent of the epidemic [5]. Because the outbreak was unanticipated, and the public health system was not prepared for an outbreak of this size, the disease spread virtually unhindered through livestock, agriculture, and the food supply. It took weeks to identify the source of the outbreak, much longer than it should have.

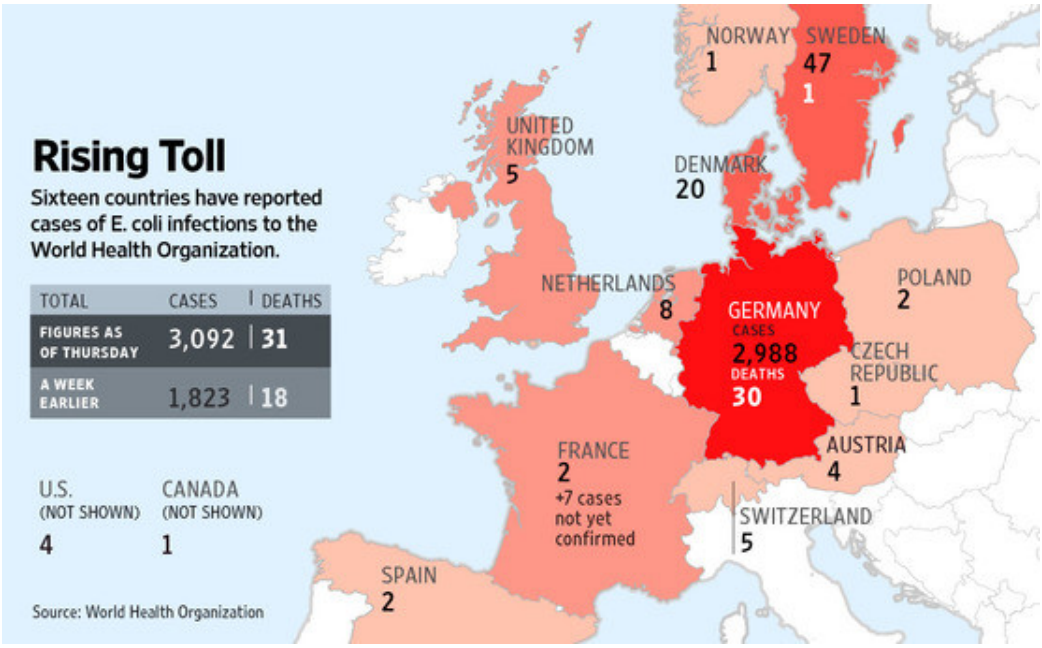


Figure 1-2: Cases of a deadly strain of *E. coli* originating in Germany and spreading throughout Europe, as of 11 June 2011. It demonstrated the need for a more effective and timely global biosurveillance capability.

This outbreak illustrates several problems still facing many biosurveillance systems. According to the National Biosurveillance Advisory Subcommittee (NBAS), this outbreak showed a lack of common terminology and a shortage of people who have the capacity to recognize and analyze these risks. It also underscored the effect of globalization on the spread of infectious disease. Another problem is the need for lateral thinking, to identify the outbreak using alternative means. Information that tracks pathogen profiles in animals and agriculture already exists in disparate

agencies, and had it been collected in a centralized database, it wouldn't have taken so long to isolate the cause of the outbreak [4]. Instead of waiting to find pathogens in the human population and then working back to find the source, effective biosurveillance should detect *E. coli* in the food supply and prevent it from reaching humans. Google Flu Trends is an example of biosurveillance that uses an alternative tracking method. It reports the incidence and spread of Influenza using certain search terms – spikes in searches for flu symptoms are a good indicator of the disease.

By cultivating the ability to quickly identify new cases of infectious disease, a society can also mitigate the economic impact of an outbreak. Various forms of biosurveillance contribute billions of dollars to U.S. spending – expenditures for hospital infection control, public health surveillance, training, research, and improvement of existing health infrastructures [1]. The implementation of a consolidated biosurveillance system could manage public health and safety more efficiently and with fewer resources.

We have thus far discussed biosurveillance in the context of the disease itself – how the disease directly impacts the population. However, the scope of biosurveillance includes the indirect consequences of the disease. Recall that biosurveillance monitors global health and disease trends to identify emerging problems; these problems are not limited to just the disease. This research focuses on the prediction and monitoring of socio-economic issues resulting from the disease spread. Specifically, we will rely on the following definition:

Negative Social Response (NSR). *A society demonstrates strain associated with the spread of an infectious disease; often results from an uncommon or unusual occurrence.*

Examples of severe behavioral negative social responses include hoarding of medical supplies, rioting, or mass flight from the region. We also consider mild non-behavioral cases like anxiety, as precursors of more unstable and hazardous situations.

Severe social responses are especially troublesome in already unstable regions, where fear can often outpace the infection. Consider recent events in Haiti –

following on the heels of a massive earthquake in January 2010, UN workers unintentionally contaminated the water supply with Cholera. It spread quickly through the population, and in two years the situation has only deteriorated as the death toll rises. As of March 2012, the NY Times reports the world has contributed $230 million to combating Haiti's unexpected epidemic, and the United Nations is now asking for an additional $53.9 million just to make it through the rainy season [6]. Many victims are seeking recompense from the UN, the epidemic has repeatedly sparked violent riots in the capital, and public mistrust of the government continues to grow. The Haitian cholera epidemic is an example of both the direct impact of a disease on the population and the costly social response that follows.

Below are several additional examples of outbreaks with social responses that negatively impacted a society.

- Bolivia, fall 2008 to spring 2009: Dengue Fever [7]
  - Most serious outbreak of the disease in the last 20 years.
  - Direct impact of the outbreak on the hospitality and tourism industry of Bolivia, and the virtual collapse of hospital and clinic systems in some areas.
- Singapore, April 2009: Gastroenteritis/Vibrio parahaemolyticus [8]
  - Market in Singapore reputed to be the center of gastroenteritis outbreak.
  - Multiple cases of fatal food poisoning, subsequently confirmed by authorities.
  - Market was temporarily closed. At the time, the Ministry of Health was engaged and the local hospitality industry was negatively impacted due to the distribution of contaminated market products.
- Argentina, May to June 2009: Influenza/respiratory disease [9][10]
  - Record absentee rates partly due to influenza/respiratory disease at schools in the greater Buenos Aires area, as well as panic buying of facemasks from

pharmacies.

– Additional reports of civil unrest during the same time period (residents attacked a bus transporting a suspected case to a local medical facility).

Although biosurveillance primarily focuses on the disease itself, we propose an in-depth examination of the *perception* of the disease, which can have equally damaging consequences. This falls under the category of quickly characterizing a bio-event to mitigate adverse health effects – once an infectious disease outbreak has been identified, the goal is to anticipate or describe the social response.

1.2 Research Objectives & Technical Approaches

The purpose of this thesis is modeling social response to the spread of an infectious disease. There are two complementary research objectives, utilizing different technical approaches:

1. Use social network analysis to describe how negative social response spreads within an outbreak.

2. Use data mining to analytically predict negative social response to an outbreak.

To accomplish these research objectives, we build our models using recent historic biosurveillance data. Social response is modeled as an idea spreading across a social network over time, and as a predictable binary response to a single outbreak. We then analyze these models using two technical approaches: social network analysis and data mining.

Social network analysis allows us to study the relationships among individuals who are represented as nodes with ties on a network [11]. This research uses agent-based modeling (ABM) to simulate interactions between individuals, and examines how those interactions spread an infectious disease and an associated social response across a complex social network. Furthermore, we test on real world data to demonstrate that the model captures the collective behavior of a society. An epidemic can be

reduced to a series of simple agent interactions on a network, which can then help describe how NSR spreads within an outbreak.

Data mining is a broad term for the process of discovering patterns in large datasets, through the use of machine learning, artificial intelligence, statistics, or a variety of other methods [12]. We employ a combination of data processing, decision trees, and classification algorithms on the biosurveillance data to predict negative social response to the spread of an infectious disease.

1.3 Thesis Organization

In Chapter 2, we describe the data and its unique characteristics, with special focus on the rarity of a true negative social response. In Chapter 3, we present the two technical approaches described above – the agent-based social network model formulation and the data-driven predictive model methodology. In Chapter 4, we report the simulation results of the social network analysis and the performance of the NSR predictions. Chapter 5 discusses our conclusions and provides recommendations for future research.

Chapter 2

The Data

In order to detect negative social response to the spread of an infectious disease, the underlying outbreak must first be detected. Biosurveillance can accomplish this in a variety of ways. Monitoring of agriculture and livestock industries can help track food-borne illnesses. Sales of over-the-counter medications can indicate unreported or undiagnosed infections, as can absenteeism from work or school. As seen with Google Flu Trends, online search terms can identify emerging Influenza outbreaks. The data used in this study takes a similar approach to identifying bio-events – rather than relying on more traditional methods like physician or hospital reporting, the data is collected primarily from media sources and the world wide web. The advantage of this approach is first and foremost near-real-time reporting of outbreaks, but the disadvantage is varying levels of reliability. While some media sources are more reliable than others, we can never assume that a report is the full and accurate truth. It is only ever the truth *as reported*. However, early warning of an outbreak is worth enduring some small uncertainty in the data. It is better to receive an early warning (true or false), than to find out too late.

This chapter will give an overview of the biosurveillance data used for this research, describe some important characteristics of the data, and conclude with two in-depth examples of real world outbreaks.

2.1 Data Overview

This study uses recent historical biosurveillance data provided by Ascel Bio, a company that leverages infectious disease forecasts and alerts to eliminate the guesswork of disease seasonality, anticipate healthcare demands, and support decision-makers [13]. The data is a historical compilation of near-real-time, multi-source biosurveillance reports, covering more than 200 countries and 30 languages. It spans over 300 infectious disease entities affecting primarily humans and animals, and was compiled using a combination of computer-based technology and human expertise. Analysts employed native linguistic and cultural context, as well as extensive professional training, to detect the earliest indicators of an infectious disease before that information filtered up through normal channels. To protect the proprietary data collection process, the following discussion of the data is limited to a high level overview.

The data is collected primarily from media sources and the world wide web, and contains approximately 11,600 labeled outbreaks of infectious disease, from September 2008 to May 2009. The available attributes in the dataset, also referred to as features or variables, provide information on the disease, the response, and the population. Each outbreak also contains information on the social response, which has been combined to create the target outcome label: "Negative Social Response" or "No Response". This is the binary classification problem we will consider further in Chapter 3.

2.2 Rarity of Negative Social Response

The most important characteristic of the Ascel Bio data is that negative social response (NSR) to the spread of an infectious disease is an exceedingly rare occurrence. Out of approximately 11,600 events, only 367 exhibit a negative social response, and very few of those are behavioral responses (such as the hoarding of medical supplies). Furthermore, there is no clear pattern of behavior to the untrained

eye that would indicate a negative social response. By looking only at the data, a non-expert cannot easily discern the difference between outbreaks that spark a negative response and those that do not. The sheer amount of no response outbreaks obscures the rare occurrence of NSR. This is a common and troublesome issue in a variety of fields, generically referred to as a "rare event detection problem". For the purposes of this research, we use the following definition:

Rare Event Detection Problem. *Less than 5% of the data are associated with indicators of negative social response.*

There are many examples of rare event detection problems outside biosurveillance, such as identifying email spam, facial detection in images, or diagnosing cancer. To better understand how to cope with a rare event problem, we examine how social response varies in more detail.

2.3 Data Characteristics

The data contains reports from 11 different regions of the world and over 200 countries. If we narrow the analysis by region, we see incredible variation in negative social response. Figure 2-1 shows that Africa (6.7%), South Asia (8.5%) and the Caribbean (9.5%) have much higher rates of NSR, despite having fewer total events than Europe (1.5%) or North America (1.6%). This suggests some kind of underlying regional differences, that either *(a)* a region is actually more prone to negatively respond to an outbreak, or *(b)* a region is more likely to report a given response as negative. Both cases indicate that region is an important factor in determining negative social response.

In addition, Figure 2-2 breaks down the composition of NSR events into mild and behavioral responses. A mild response is typically characterized as anxiety about the spreading infection. It can also occur when media sources report panic in the population, but there is no evidence of a behavioral response. Note the variation in the level of behavioral cases; for example, of all Caribbean outbreaks with negative

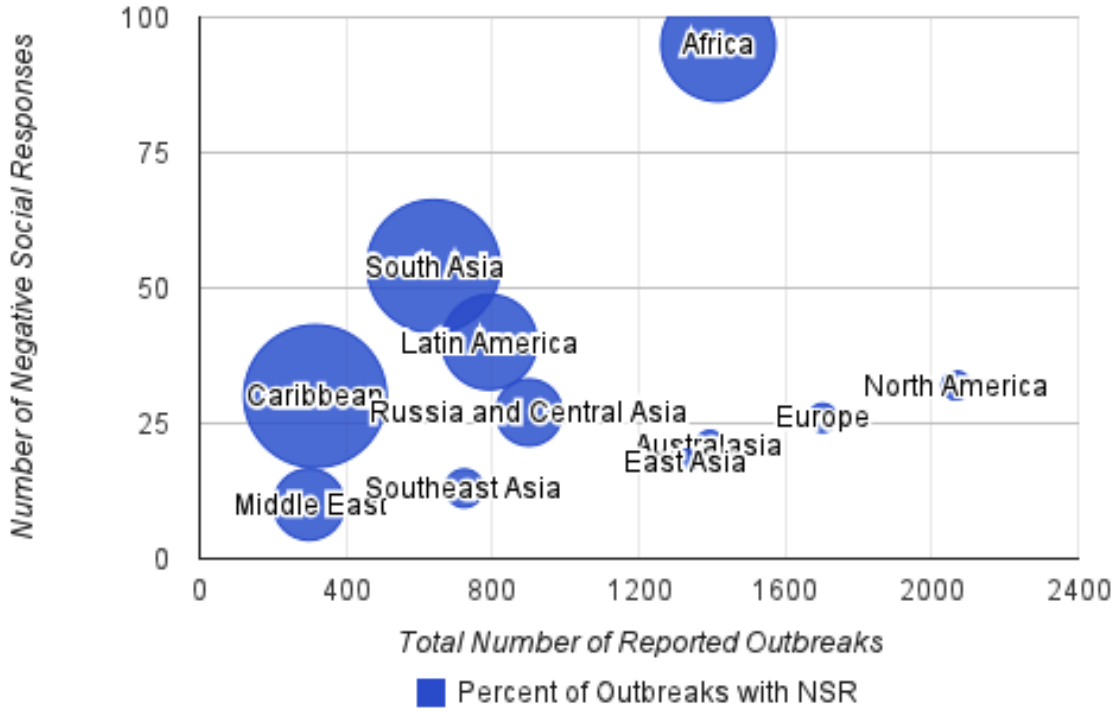


Figure 2-1: Negative Social Response vs. Number of Outbreaks. Incidence rate of NSR varies widely by region.

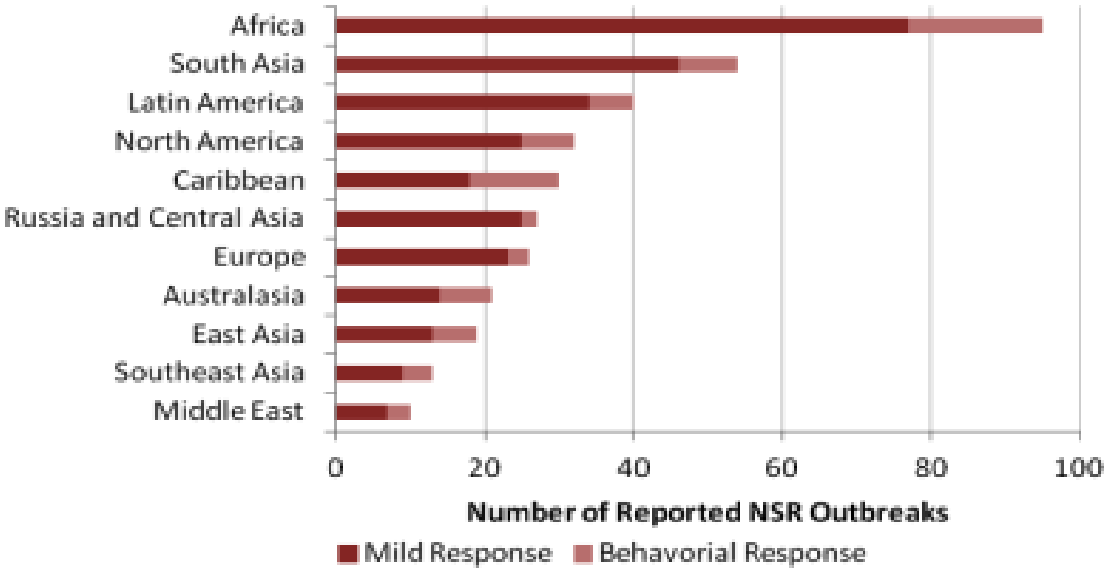


Figure 2-2: Region vs. Number of NSR Outbreaks. Composition of NSR events varies widely by region.

social responses, 40% are behavioral, while only 7% are behavioral in Russia and Central Asia. Not only are some regions more likely to exhibit negative social response to the spread of an infectious disease, but those responses are also more likely to be extreme.

The data can also be examined by disease. There are over 170 diseases reported in the data; some are specific infections (e.g. Plague, caused by the bacterium *Yersinia pestis*), while others are generic diseases with multiple underlying causes (e.g. Respiratory Disease or Gastroenteritis). If we narrow the analysis by disease, we again see a large amount of variation in negative social response.

The 10 most commonly reported diseases in the data are shown in Figure 2-3. Although Chikungunya Fever (CHIKV) has roughly the same number of outbreaks as Salmonella Infection, CHIKV is almost ten times more likely to inspire a negative social response. One reason for this difference is that while Salmonella is a common bacterial disease affecting the intestinal tract and is usually not serious for healthy individuals, CHIKV is a painful and increasingly widespread disease. Additionally,

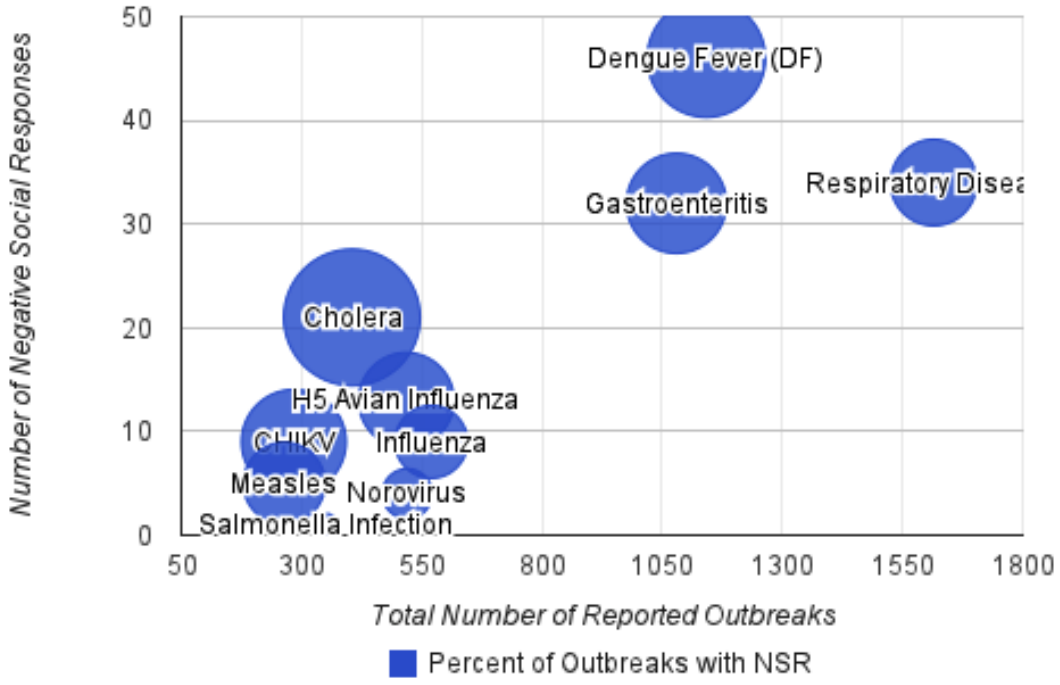


Figure 2-3: Negative Social Response vs. Number of Outbreaks. Top 10 most common diseases show the incidence rate of NSR varies widely by disease.

during the period represented in this data, CHIKV was spreading eastward from Africa with unusually severe symptoms. Inhabitants mistook the disease for epidemic Dengue Fever or Dengue Hemorrhagic Fever, generating a negative social response. This example illustrates why the data shows variation in NSR by disease – much depends on the context of pathogen presentation within a community.

In addition to seeing social responses vary by disease, we also see common diseases vary by region. NSR highly depends on the nature of the infection, the location of the outbreak, and whether the infection is endemic. In epidemiology, an infection is said to be endemic when it is maintained in the population without the need for external inputs. While there are many other factors that drive negative social response, disease and region contribute the majority of influence. As an example, compare North America to South Asia. Figure 2-4 illustrates the extent of the differences between two regions. Of the ten most common diseases in South Asia and the ten most common diseases in North America, only three are shared by both – Respiratory Disease, Gastroenteritis, and Dengue Fever. For South Asia, diseases like Malaria

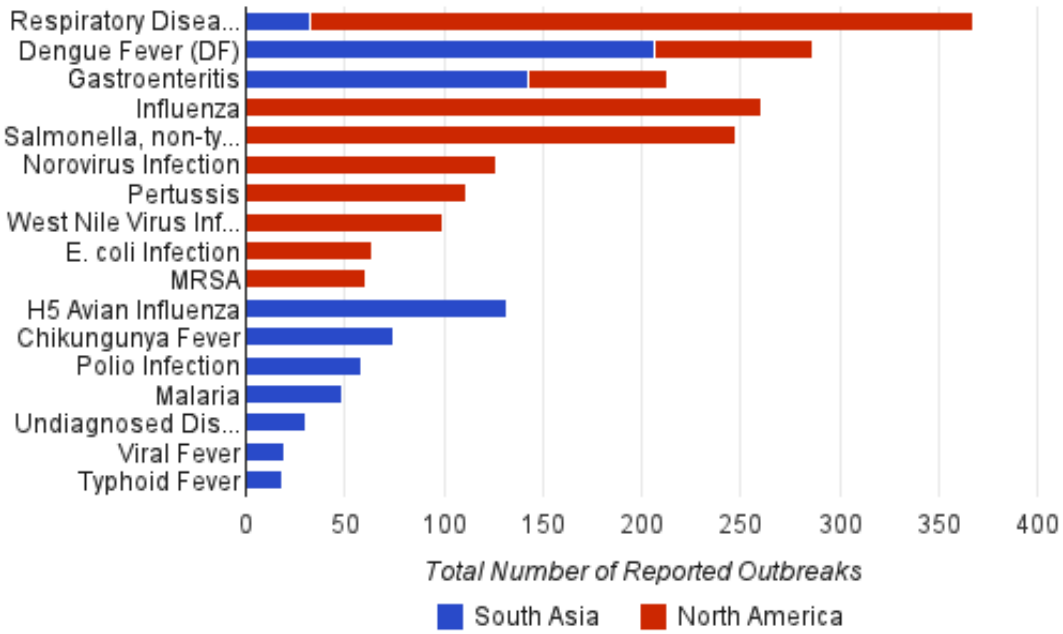


Figure 2-4: Comparison of the most common diseases in South Asia and North America, showing the wide disparity between regions. Only 3 diseases are common enough to be represented in the top 10 of *both* regions.

and Typhoid Fever are routine, while on the other side of the world, Influenza is common.

What happens when the situation changes? West Nile Virus originated in Africa and was first discovered in the United States in 1999. By 2002, thousands of individuals became infected and the CDC recorded several hundred fatalities [14]. The outbreak strained resources and spread anxiety about how to combat the then relatively unknown disease. It is now endemic to the US. The spread of West Nile Virus is a classic example of how multiple factors combine to drive public perception.

Although there are many factors that precipitate negative social response, often the disease itself and how it presents in the community is the primary impetus. Figure 2-5 on the following page displays the most commonly reported infections in the data. Negative social response is generally too rare to visually determine any variation, but numerically the incidence rate of NSR for the top diseases ranges from zero in many cases to 17% with Anthrax (for good reason).

The data also contains many other variables that influence social response to the spread of an infectious disease. Since they cannot be discussed for contractual reasons, this study presents region and disease as important illustrations of the complex nature of NSR. We propose that the data accurately captures the characteristics of real world bio-events, and we will use this information to help model social response to the spread of an infectious disease.

2.4 Temporal Features of the Data

In addition to the 11,600 initial outbreaks reported in the data, there are 9,200 reports that contain updated outbreak information. An initial bio-event may have no updates, it may have several, or in the case of an epidemic there may be 100 updates. Thus, for certain outbreaks we can establish a timeline of how the infection progressed. While our research only considers the initial outbreaks for the data-driven predictive analysis, we utilize updates to establish the spread of the disease for an individual outbreak, and apply this information to the social network analysis.

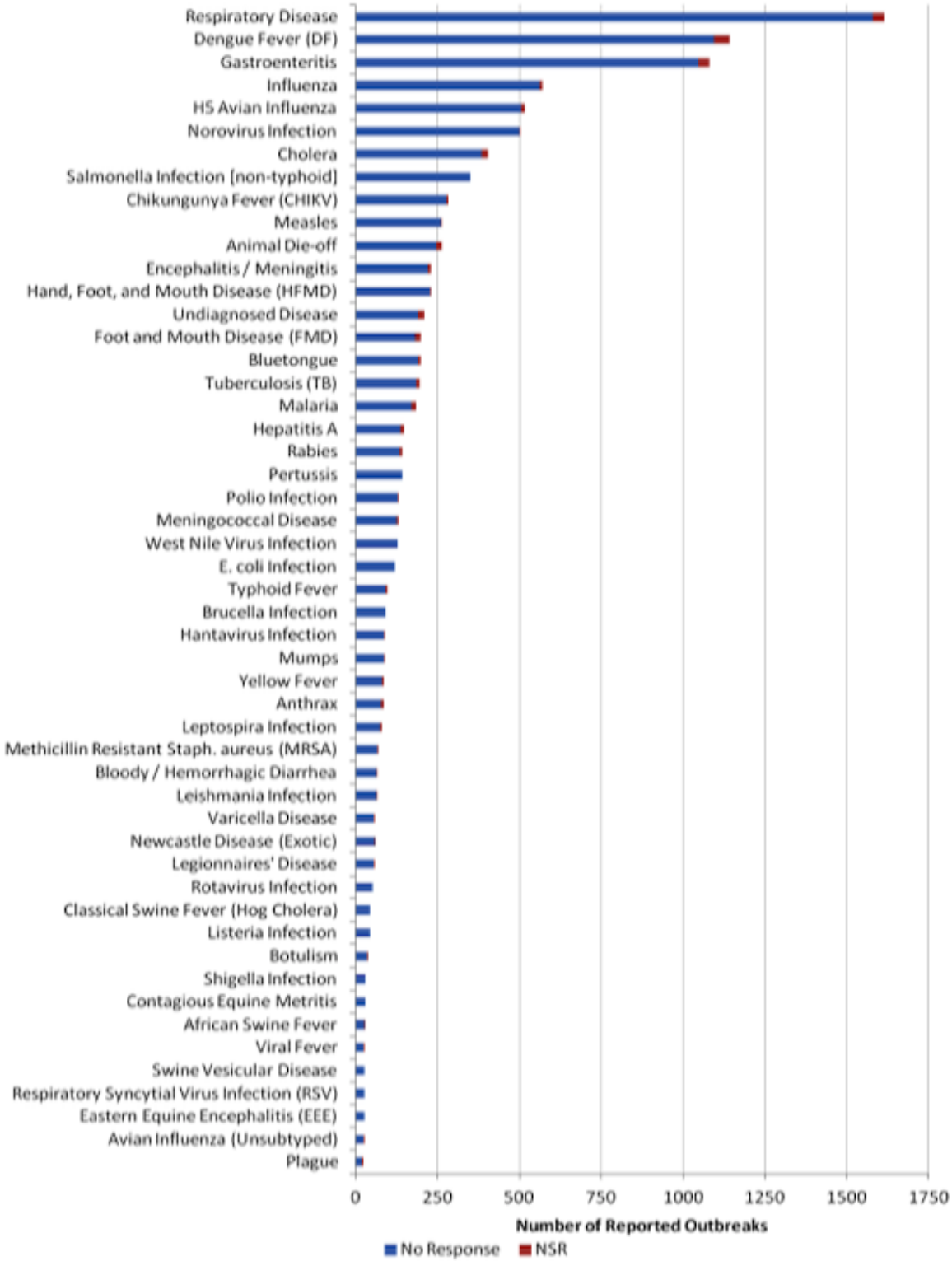


Figure 2-5: Disease vs. Number of Outbreaks. Top 50 most common diseases. Incidence rate of NSR varies from 0-17%.

2.5 Example Outbreaks

Two outbreaks of Dengue Fever were selected from the data to contrast the difference between an outbreak with no response and one with an exaggerated negative social response. Although the specific locations of the outbreaks have been redacted, they are both real world events that took place in large cities. City X was selected for its lack of social response, and City Y was selected for more extreme NSR. These examples were also chosen because of the amount of information available in the updates. After the initial report of the outbreak, updated reports include the number of individuals infected, changes in the outbreak, and the social response. In addition, Outbreak X and Outbreak Y will later be used to validate the social network model in the results.

2.5.1 City X in South Asia

From September to October 2008, a large urban metropolis in South Asia experienced a typical outbreak of Dengue Fever. The disease is endemic to the region, as it is in most tropical areas of the world that experience long rainy seasons. Figure 2-6 and Table 2.1 show the progression of the disease, from the first reports to when new infections began to decline, as reported in the biosurveillance data.

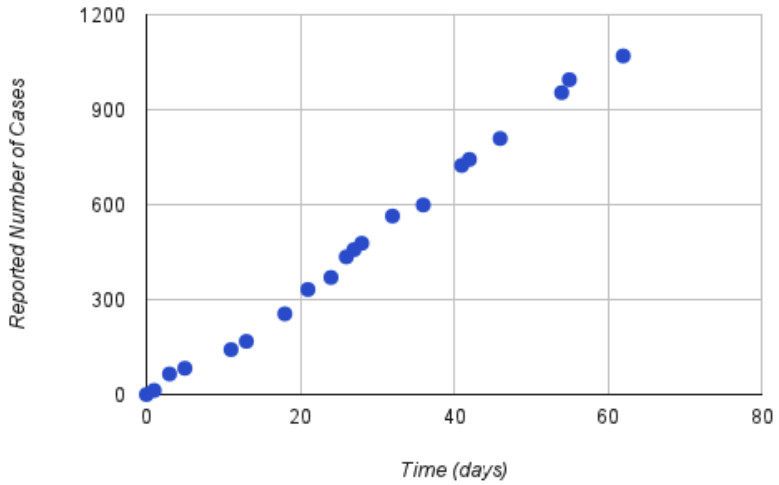


Figure 2-6: City X in South Asia: total number of reported cases over time.

Table 2.1: Detailed Updates of Dengue Fever event in City X

Day	NSR	Detailed Reports
1	No	13 new confirmed cases of Dengue with 1 fatality. Normal outbreak for this time of year.
3	No	65 total cases.
5	No	83 total cases.
11	Yes	142 total cases; sudden spurt of new infections.
13	No	168 total cases; local authorities predict that the situation will be brought under control within ten days.
18	No	255 total cases.
21	No	47 new cases of Dengue; another fatality.
24	No	370 total cases.
26	No	435 total cases; government reports no need to panic, the number of Dengue cases are not higher than expected.
27	No	458 total cases.
28	No	478 total cases; approaching epidemic proportions.
32	No	564 total cases; contrary reports state that outbreak is both normal and higher than expected.
36	No	599 total cases.
41	No	724 total cases.
42	No	743 total cases.
46	No	809 total cases.
54	No	954 total cases.
55	No	995 total cases; increase in cases this year likely due to changing weather conditions.
62	No	1070 total cases, most confirmed; declining number of new infections due to dropping temperatures and pollution.

City X is an example of an infectious disease that exhibited no serious social response. This outbreak occurred in a city that routinely experiences Dengue Fever and has sufficient public health infrastructure to handle the extent of the infections. New cases confirmed by medical professionals, and multiple media sources corroborate the number of infections. This is a very reliable sequence of reports documenting the spread of a typical Dengue Fever outbreak in South Asia. We will use it as a type of baseline outbreak to contrast the differences between an outbreak with no response and one with severe NSR.

2.5.2 City Y in Latin America

In March 2008, a large city in Latin America experienced an atypical outbreak of Dengue Fever. The disease is endemic to the region, but highly dependent on the severity of the rainy season. Figure 2-7 and Table 2.2 show the progression of the disease, from the first reports to when new infections began to decline, as reported in the biosurveillance data. Note that the total number of reported cases shows exponential growth; often, the curve of social reporting mirrors forensically reconstructed epidemic curves.

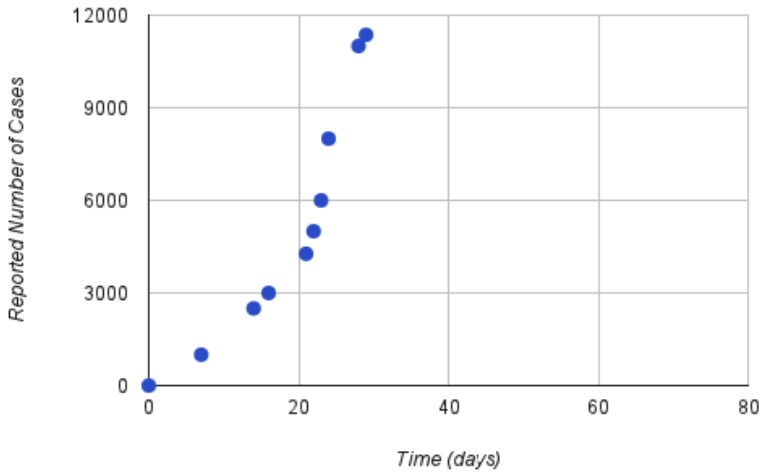


Figure 2-7: City Y in Latin America: total number of reported cases over time. Despite exaggerated numbers, we see a realistic progression of the disease.

City Y is an example of an infectious disease that sparked a widespread negative social response. Furthermore, it was notable for the contradictory reports of what was happening in the region, suggesting that the reported number of cases is likely not the true spread of the disease. Although the outbreak in CIty X resulted in approximately 1,000 cases, and the outbreak in City Y generated reports of 10,000+ cases, the increase in reported infections is likely *not the cause* of the NSR; rather, it is a *symptom* of the negative social response. As people become increasingly agitated in response to the disease, reports become exaggerated and spread through word-of-mouth. By the end of the outbreak, only one-tenth of the reported cases were actually confirmed by medical professionals. Often, confirmation of cases significantly

lags behind real-time reporting. We must assume that the progression of the disease is relatively accurate but may be exaggerated in scale, and that "confirmation by authorities" is actually "confirmation of the sample size the organization in question was able to acquire."

Table 2.2: Detailed Updates of Dengue Fever event in City Y

Day	NSR	Detailed Reports
7	Yes	Contrary Dengue Fever statistics reported; 1000 total cases, with 135 suspected officially and 9 confirmed.
14	Yes	2500 total cases; 1 suspected fatality; reports of hospitals flooded with people seeking treatment.
16	Yes	3000 total cases; 3 suspected fatalities; anxiety among residents; hospital collapse due to elevated patient levels; Mayor worried and national authorities become involved.
21	Yes	4270 total cases; residents alarmed, situation desperate, increasing public health strain; situation exacerbated by rivalries between local and national authorities.
22	Yes	5000 total cases; reports of overflowing hospitals and collective psychosis; residents panicked but "everything under control".
23	Yes	6000 total cases; chaos due to politicization of situation; some officials claiming there is no epidemic; reports of families fleeing the city.
24	Yes	8000 total cases; lack of medical supplies, continued denial by public officials of epidemic status; only 1/10 the number of cases have been confirmed while other aid organizations are reported ever higher numbers of people infected.
28	Yes	11000 total cases; reported infection rates continue to vary.
29	Yes	11363 total cases; 2 fatalities; final estimates of provincial Dengue cases range from 1000 confirmed to 11000 reported.

The outbreak in City Y is also noticeably shorter in duration that that of City X. We infer that the steep increase in new infections contributed to the NSR, as indicated by the reports of hospital collapse and public health strain within a week of the initial infections. While we can make these types of inferences about the social response, the data does not describe specific numbers about the level of NSR or how quickly it spreads. The models presented in the following chapters will help provide some insight into this issue.

Chapter 3

Models & Methodology

In this thesis, we build two models to describe social response to the spread of an infectious disease. The models utilize the biosurveillance data in different ways, to predict negative response to an outbreak and to understand how that response spreads within the outbreak. After giving an overview of the models, this chapter provides the two model formulations in more detail.

The *Epidemic Social Response Model* is based on a social network analysis. It builds a network of individuals connected through social ties and analyzes how the interactions of agents spread infection and negative social response. To validate the model, we test it on real world examples from the data (Dengue Fever events in City X and City Y). The model simulation shows that the complex behavior of an outbreak can be successfully explained by a series of agent interactions on a network. This provides useful insight into the mechanism that drives the spread of social response to an outbreak.

The *Data-driven Predictive Model* uses supervised learning and a data mining approach to data. It analytically predict outbreaks that result in negative social response. The model employed a joint methodology to improve performance, developed in the process of our research, that uses additional interaction features, data enrichment, and a voting strategy. Interaction features and data enrichment help the learning algorithms to better identify outbreaks with no response, by increasing model specificity and creating a dataset where NSR is less rare. The voting strategy

combines the positive characteristics of two learning algorithms, Logistic Regression and Random Forests. The joint methodology as a whole improves performance through the reduction of false positives, yielding reasonably accurate predictions of negative social response to the spread of an infectious disease.

The overview below of the modeling approach illustrates the interconnected nature of the disease and the social response.

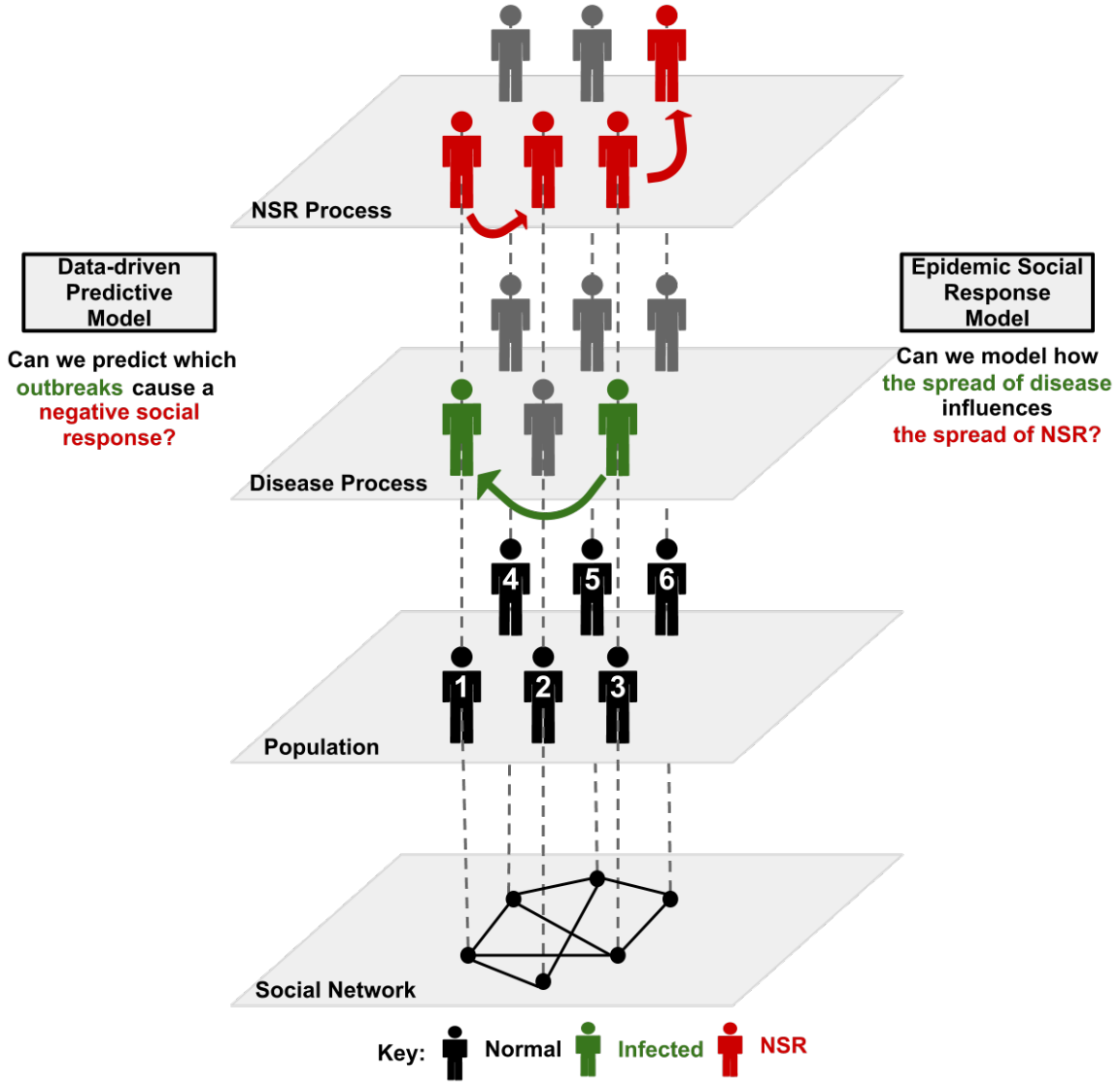


Figure 3-1: Modeling approach – the two models represent complementary views of social response to the spread of an infectious disease. The *Epidemic Social Response Model* takes a micro view and analyzes how disease and NSR spread within an outbreak. The *Data-driven Predictive Model* takes a macro view and analytically predicts negative social response to the outbreak as a whole.

3.1 *Epidemic Social Response Model*

3.1.1 Modeling Approach

We aim to model negative social response to the spread of an infectious disease using social network analysis. Because NSR is rare and does not show clear patterns of behavior, it is not immediately clear from the data why one outbreak sparks a negative response and another similar outbreak does not. The *Epidemic Social Response Model* simulates an outbreak and the associated strain on a population to better understand how different outbreaks influence social response.

The *Epidemic Social Response Model* uses a simulation technique called Agent-Based Modeling. ABM models a real world system as a collection of autonomous decision-making agents, whose decisions are based on a set of rules [15]. We apply ABM to a social network, where individuals in a population are connected to others through social ties. These agents then repeatedly interact with each other according to simple behavioral rules, but generate complex behavior. Emergent phenomena such as this occur when a co-operation of unlike things cannot be reduced to their sum or their difference; the whole is more than the sum of its parts [16]. The advantage of ABM is its effectiveness at capturing emergent phenomena, which makes it ideal for modeling social response. NSR is a classic example of emergence – when a population first begins to show strain, the collective behavior of the crowd quickly eclipses that of any one individual.

Agent-Based Modeling provides the framework for the social network analysis. Each agent in the network has a condition consisting of his disease class (Susceptible, Infected, or Recovered) and his negative response level (between 0 and 10). As agents interact within the network, their conditions change according to a set of probabilistic rules, forming two interconnected processes: the spread of negative social response "on top of" the spread of disease, as shown in Figure 3-1. These processes can be considered almost entirely separate from each other except for a simple coupling device – when an agent becomes infected, his negative response level is maximized. In the absence of this situation occurring, each process proceeds according to its own

set of condition dynamics.

The coupling device creates an overall *Epidemic Social Response Model*, and allows us to study how the spread of disease influences social response. The following sections will examine the relevant literature and present the mathematical model formulation.

3.1.2 Related Literature

One of the primary reasons for studying networks is to comprehend the mechanisms by which things like information, innovations, or disease spread over them. For example, the main reason for the study of sexual contact networks is to help us understand and potentially limit the spread of sexually transmitted diseases [17]. Social network analysis is a broad term and includes the study of network structure, network generation, and how something spreads over a network.

Although there are many models relating to the spread of infection over a network, or the spread of information over a network, models for the interaction of these processes is comparatively limited. However, the work of Meloni et al [18] is very similar to this thesis in subject matter. Their model of human mobility responses to the large-scale spreading of infectious diseases examined how epidemic spread induces self-initiated changes in behavior, which in turn influences the spread of the disease. For instance, in response to an outbreak, individuals may decide to flee, inadvertently spreading the disease. To describe this feedback loop, they use a meta-population approach, which is a framework that describes a set of subpopulations as a network whose links denote individual mobility across subpopulations. Meloni concludes that the real-time availability of information on the disease and on how people react to the disease can have a negative impact on disease containment and mitigation – self-initiated behavioral changes may enhance disease spread [18].

There are several epidemiological models for the spread of disease over a network. The simplest is the SIS model, or susceptible-infected-susceptible model. However, an expanded and more widely used model is SIR, first proposed in 1927 by Kermack and McKendrick [19]. It separates the population into three classes: susceptible (S), meaning they aren't infected but can catch it if exposed to someone who is, infective

(I) meaning they have the disease and can transmit it to others, and recovered (R), meaning they have recovered from the disease and are permanently immune, so that they can never catch it again or pass it on to others [17]. In standard mathematical epidemiology it is assumed that any susceptible individual has a uniform probability β per unit time of becoming infected and that infective individuals recover and become immune at some stochastically constant rate γ. The fractions s, i and r of individuals in the states S, I and R are then controlled by the differential equations:

$$\frac{ds}{dt} = -\beta is, \frac{di}{dt} = \beta is - \gamma i, \frac{dr}{dt} = \gamma i$$

This is the basic SIR formulation of an infection spread over time; as we will see in this thesis, it can be generalized to fit a variety of situations. Other models, shown in Figure 3-2, add classes to represent specific disease dynamics [20]. For example, babies are often not immediately susceptible and are born into a separate class, due to a period of maternally-derived immunity to a disease (such as measles).

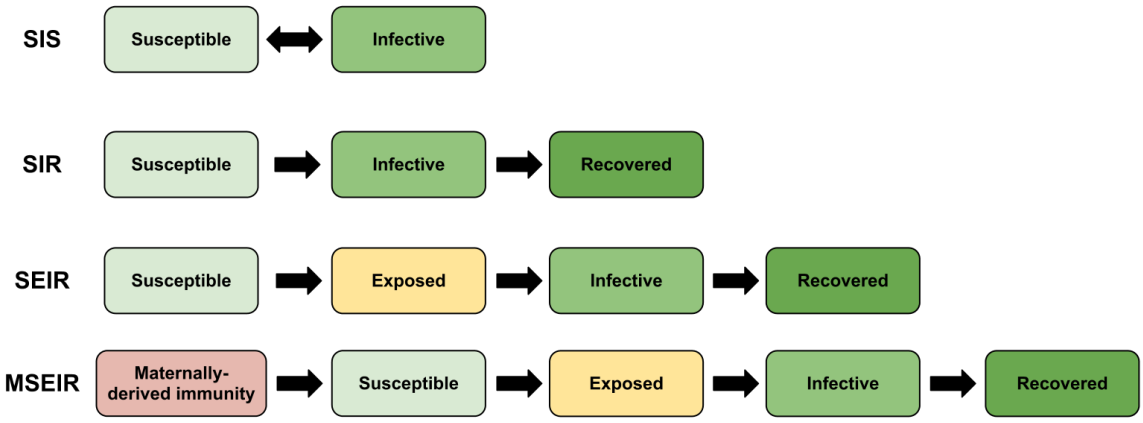


Figure 3-2: Common epidemiological models show wide variation to adapt to different diseases and populations

In addition to studying how disease spreads, social network analysis can also be used to study how an idea or belief spreads among individuals, referred to as opinion (or attitude) dynamics. Katz and Lazarsfeld [21] first posited that a small number of individuals play a critical role in shaping public opinion. They developed a model of communication to explain the diffusion of ideas, innovations, and commercial

products; their concept was that ideas spread from media to opinion leaders, and then to their primary social groups. These influential people were the key to a wide diffusion of ideas.

Abelson [22] took the next step towards formulating a mathematical model to characterize how a pair-wise interaction of people can affect their scalar-value attitudes. The "persuasiveness" of each individual and the difference in their attitudes generates this change in beliefs and can be written as a system of differential equations. He also noted a limitation of his model, that every individual reaches universal agreement. A decade later, de Groot [23] used the theory of Markov chains to better model the weight an individual gives to the opinion of his neighbors. His work helped lay the foundation for future study in the field, especially the model most critical to this thesis.

The spread of misinformation model developed by Acemoglu et al [24] characterized agent beliefs in a social network, given varying levels of influence among agents. The spread of misinformation model posited that in a network, pair-wise interactions between agents were probabilistic in the frequency of their meeting, and in the type of interaction between the agents. Assuming that every agent in the network is influenced by someone else ("no man is an island"), this work demonstrated that the presence of more influential forceful agents leads to the formation of a consensus [24]. This model can be applied to almost any type of information being spread over a social network, and thus provides a simple and relevant basis for the dynamics of the *Epidemic Social Response Model*.

3.1.3 Network Topology

While there is a large component of social network research devoted to network structure and generation, this study relies on a random network of individuals connected through social ties. The model is validated using real world examples from the data, and in the absence of information regarding societal structure, uses a generic representation of the population. We implemented a common random network topology generator first proposed by Waxman [25]. Model simulations are performed

on the undirected graph shown in Figure 3-3.

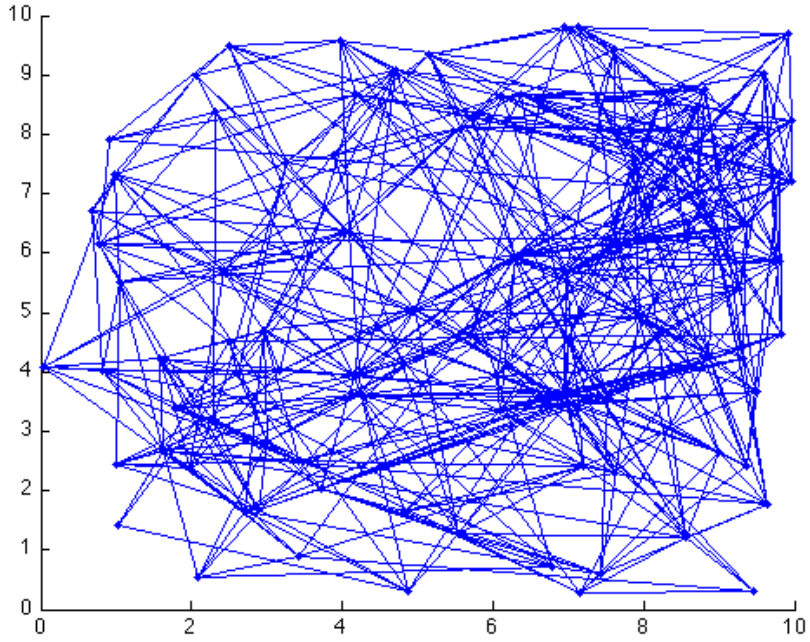


Figure 3-3: Two-dimensional representation of 100-node random test network used for simulations

There are several important observations about the network topology. First, the model is scalable. The size of the test network shown here is scaled down to allow analysis of much larger populations. Second, there is a wide variety of network topologies, and simulations can sometimes perform very differently on alternate structures (e.g. line, ring, or tree networks). The Waxman random graph, $G(N, E)$, is chosen for its generality and flexibility. Third, we assume the same contact network for both the disease and NSR process, which may not be the case in the real world. Future work will expand the analysis of the model to other network topologies, and differentiate between social ties that spread infection, and social ties that spread ideas.

3.1.4 Agent Interactions

The basic model characteristics and dynamics are based on the spread of misinformation model from Acemoglu, et al [24]. Interactions between agents occur in a pair-wise manner according to a Poisson process with rate 1, independent of all other

agents. In a network of n individuals, interactions thus occur as a Poisson process with rate n. A consequence of assuming a Poisson process is that there is at most one interaction at any given time, and these interactions are indexed over all agents with k, $k \geq 1$. Note that this implies the time between interactions follows an exponential distribution and is not fixed.

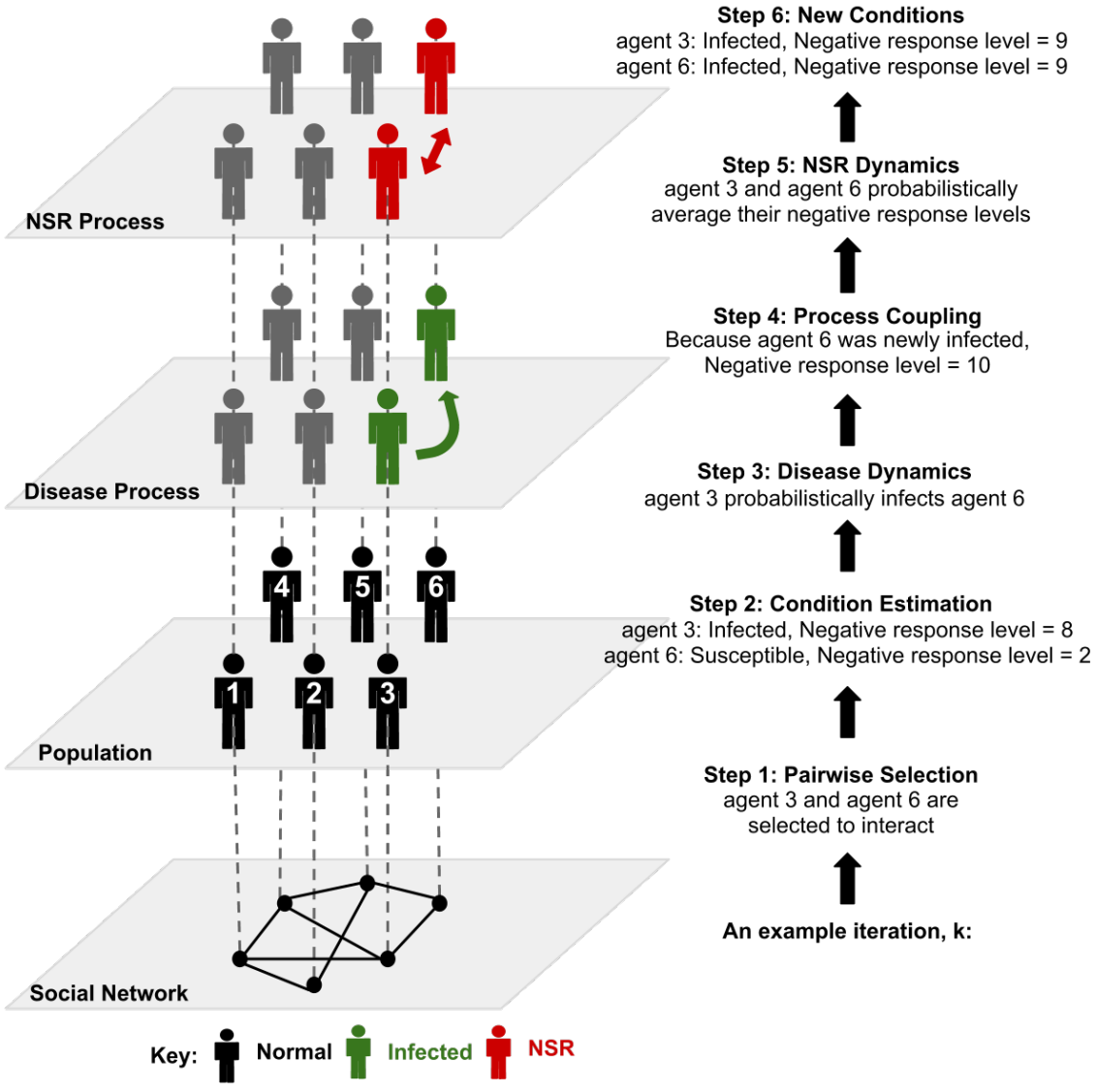


Figure 3-4: A typical iteration of the model, where one agent infects the other and both agents average their negative response levels

At each interaction k, a series of events occurs, as illustrated in Figure 3-4. Each agent in the network has a condition consisting of his disease class (Susceptible, Infected, or Recovered) and his negative response level (between 0 and 10). As agents interact within the network, their conditions probabilistically change according to a set of rules, forming two interconnected processes: the spread of negative social response on top of the spread of disease. They can be thought of as two distinct but parallel processes that have been coupled to form one larger model, based on the same underlying model foundation. We chose to formulate the model as two connected processes to differentiate their effects – we can individually examine how disease spreads, how social response spreads, and how disease influences social response.

The main components of the agent interaction process will be discussed further in the following sections: condition estimation, condition dynamics, and process coupling.

3.1.5 Condition Estimation

At each interaction k, an agent's condition consists of both his disease class and his negative response level.

Disease Estimation

An agent's condition can be estimated as a discrete random variable, $D_i(k) \in \{S, I, R\}$, where we denote $D_i(k)$ as agent i's disease class at the k-th interaction. The interpretation is that at any given interaction, an agent can be either susceptible, infected, or recovered.

Figure 3-5: Disease Process, discrete classes

NSR Estimation

An agent's condition can be estimated as a continuous random variable, $P_i(k) \in [0, 10]$, where we denote $P_i(k)$ as agent i's NSR level at the k-th interaction. The interpretation is that at any given interaction, an agent's negative response level ranges from 0 to 10, inclusive.

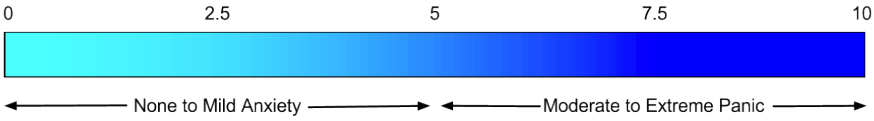


Figure 3-6: NSR Process, continuous negative response scale

3.1.6 Condition Dynamics

Next we model the condition dynamics of all agents as a Markov chain. The state of the system at any point in time is the set of agent conditions at interaction k, $P_i(k)\ \forall i \in N$ and $D_i(k)\ \forall i \in N$. The probabilistic pair-wise interactions determine the state transitions. We assume the Markov property holds, that given the current state, the state will have the same transition probabilities to another state, regardless of past transitions [26]. In particular, agents change conditions as a result of memoryless pair-wise interactions with neighbors in the network.

The *Epidemic Social Response Model* uses the same system of agent interactions and dynamics as Acemoglu's spread of misinformation model, which is concerned with how attitudes spread across a network. Given a set of agent conditions, the disease process and the NSR process make the following modifications:

- In Acemoglu's model [24], an agent's attitude can probabilistically change in three ways (forceful, averaging, or identity).
  - In the NSR process, agent conditions can probabilistically change in four ways (forceful, averaging, decay, or identity).
  - In the disease process, agent conditions can probabilistically change in three ways (infection, recovery, or identity).

- Acemoglu's model is a single process.
  - The NSR process and the disease process are each updated from the spread of misinformation model. The two processes together are coupled to form a larger model.

Disease dynamics

At each interaction k:

- Agent i initiates an interaction according to a uniform probability distribution. Agent i then selects agent j uniformly at random from his neighbors.
- Conditioned on agents i and j meeting, the following pair-wise interactions can occur:
  1. *Infection*. With probability ν_{ij}, agent i infects agent j:

$$D_i(k+1) = D_j(k+1) = D_i(k) \tag{3.1}$$

  2. *Recovery*. With probability κ_{ij}, agent i recovers:

$$\begin{aligned} D_i(k+1) &= R \\ D_j(k+1) &= D_j(k) \end{aligned} \tag{3.2}$$

  3. *Identity*. With probability $(1 - \nu_{ij} - \kappa_{ij})$, nothing changes:

$$\begin{aligned} D_i(k+1) &= D_i(k) \\ D_j(k+1) &= D_j(k) \end{aligned} \tag{3.3}$$

- The interaction probabilities are subject to the following rule, where ν is the fixed probability of infection and κ is the fixed probability of recovery:

$$\nu_{ij}, \kappa_{ij} = \begin{cases} \nu, \kappa & \text{if } D_i(k) = I; \\ 0 & \text{else.} \end{cases} \tag{3.4}$$

The interpretations of the disease interactions are straightforward. Conditioned on two agents interacting, there are only two cases where a state transition occurs. Either an infected agent infects a susceptible agent (Interaction 3.1), or an infected agent recovers (Interaction 3.2). In all other cases, e.g. both susceptible or both recovered, there is no change in $D_i(k)\ \forall i \in N$.

NSR dynamics

At each interaction k:

- Agent i initiates an interaction according to a uniform probability distribution. Agent i then selects agent j uniformly at random from his neighbors.
- Conditioned on agents i and j meeting, the following pair-wise interactions can occur:

    1. *Forceful.* With probability α_{ij}, agent i forcefully influences agent j's NSR level:

    $$P_i(k+1) = P_j(k+1) = P_i(k) \tag{3.5}$$

    2. *Averaging.* With fixed probability β, agents i and j average their NSR levels:

    $$P_i(k+1) = P_j(k+1) = \frac{P_i(k) + P_j(k)}{2} \tag{3.6}$$

    3. *Decay.* With fixed probability δ, both agent's NSR level decreases by a fixed decay parameter, Δ:

    $$\begin{aligned} P_i(k+1) &= \Delta * P_i(k) \\ P_j(k+1) &= \Delta * P_j(k) \end{aligned} \tag{3.7}$$

    4. *Identity.* With probability $(1 - \alpha_{ij} - \beta - \delta)$, nothing changes:

    $$\begin{aligned} P_i(k+1) &= P_i(k) \\ P_j(k+1) &= P_j(k) \end{aligned} \tag{3.8}$$

- The interaction probabilities are subject to the following rule, where α_{HI} and α_{LO} are fixed probabilities:

$$\alpha_{ij} = \begin{cases} \alpha_{HI} & \text{if } P_i(k) > 5; \\ \alpha_{LO} & \text{if } P_i(k) \leq 5. \end{cases} \tag{3.9}$$

The interpretations of the negatives social response interactions are more complex. Conditioned on two agents interacting, there are several ways an agent's NSR level can change. The forceful interaction is the most important, as it captures the idea that *NSR is itself contagious*. If $\alpha_{HI} = \alpha_{LO}$, this implies that an agent is just as likely to force his extreme panic on another agent as he is to force his extreme calm. However, if $\alpha_{HI} > \alpha_{LO}$, the interpretation is that the negative response will spread quickly when agents are forced to become extremely agitated. Interaction 3.6, on the other hand, is a more common measure of attitude exchange – it represents both agents reaching a consensus about their social response to the situation. The decay parameter (Interaction 3.7) can be interpreted as a natural "calming down" effect over a long period of time. If none of these interactions occur, agents retain their NSR level and there is no change in $P_i(k)$ $\forall i \in N$.

Disease and NSR Coupling

Coupling is a technique that probabilistically joins two Markov processes together [27]. The coupling generates an overall combined process that may or may not by Markovian itself, which is beyond the scope of this research. The binding of the NSR process to the disease process is the final addition to the model: when an agent first becomes infected, he also develops an extreme negative social response. In mathematical terms,

$$\text{If } D_i(k) = S \text{ and } D_i(k+1) = I, \text{ then } P_i(k+1) = 10, \forall i \in N. \tag{3.10}$$

The *Epidemic Social Response Model* uses this rudimentary coupling device to individually examine how disease influences social response. As agents interact on the network, the disease process and the NSR process intertwine to produce complex emergent behavior, which can help explain the mechanism that drives negative social response to the spread of an infectious disease.

3.2 *Data-Driven Predictive Model*

3.2.1 Modeling Approach

We also aim to model negative social response to the spread of an infectious disease using a data mining approach. The *Data-driven Predictive Model* uses supervised learning to analytically predict outbreaks that result in negative social response. This type of statistical learning uses feature values to find patterns and relationships between these features and the outcome. Although supervised predictive models work well in situations where historical data contains many examples of the outcome, in a situation with rare event occurrence, prediction is more difficult. We have developed a methodology that specifically targets the rare event detection problem. Our joint methodology, shown in Figure 3-7, improves the performance of the model over the baseline methods. It is a sequential process of creating interaction features to the data, enriching the data, and applying a voting strategy to multiple classifiers.

The baseline methodology builds a supervised learning model on the labeled data to predict the known target outcome value (NSR). We will show in Chapter 4 that when applied to the biosurveillance data, Logistic Regression results in a high number of false positives, while Random Forests results in a high number of false negatives. The creation of interaction features and data enrichment help the learning algorithms to better identify outbreaks with no response, by increasing model specificity and creating a dataset where NSR is less rare. The voting strategy combines the complementary nature of the learning algorithms to boost identification of true NSR events. The joint methodology as a whole significantly improves performance,

yielding reasonably accurate predictions of negative social response to the spread of an infectious disease. The following sections will introduce some background information and the underlying methods employed, followed by the overall predictive model.

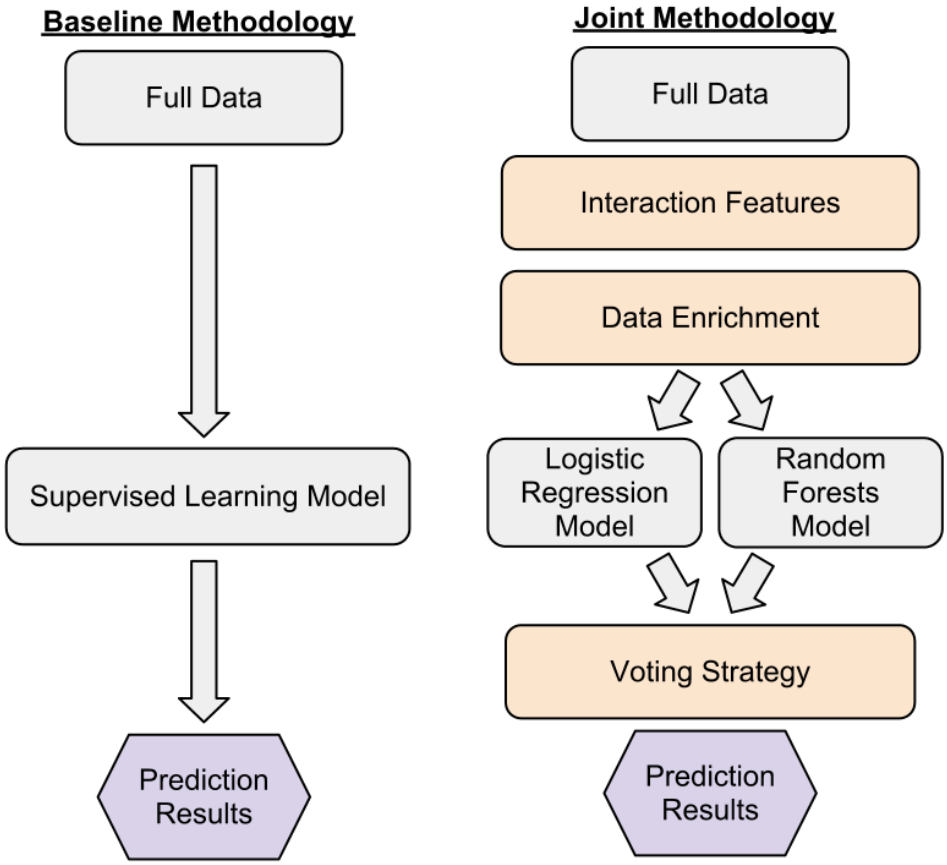


Figure 3-7: Overview of the *Data-driven Predictive Model*, which utilizes the joint methodology to improve upon the baseline methods

3.2.2 Related Literature

The literature most relevant to the *Data-driven Predictive Model* informally applies the basic idea of ensemble learning: can we combine multiple classifiers into a single strong classifier that improves upon the performance of any one individual algorithm? Both the data enrichment and the voting strategy used in the joint methodology of this research are derived from this idea.

Basic probability theory posits that all evidence relevant to a prediction should be used in making that prediction [28]. This implies that if a new prediction method

provides more relevant information, then it can be combined with existing methods to improve predictability. This idea is the foundation of ensemble learning, which builds multiple classifiers for a single prediction task. A popular type of ensemble learning is bagging, where different prediction models are repeatedly trained using different bootstrap samples of the data, and then combined using some kind of voting method[12]. Although bagging generally relies on fast, simple learning algorithms to create an abundance of models, a voting strategy can be applied to almost any set of classifiers.

Carpenter et al [29] developed a voting strategy for neural networks that was later applied by Downs et al [30] to diagnose breast cancer from fine-needle aspirates of the breast. They built many individual neural networks and tested the five most accurate using a voting strategy. Each network makes a prediction for a test record, the number of predictions made for each category (benign or malignant) is totaled, and the category with the most number of votes is the final predicted outcome. Downs discovered that although the voting strategy only made a slight improvement in accuracy over the individual networks, it provided a useful partitioning between outcomes with low certainty and high certainty [30].

Another popular type of ensemble learning is boosting, based on the question posed by Kearns [31]: can a set of weak learners create a single strong learner? In rare event detection problems, the basic idea of boosting informs a variety of methods. Wu et al [32] uses a cascade learning method for facial detection in images. They apply a stage-wise, greedy feature selection process that considers different regions of an image, and rejects those that clearly do not contain faces. The algorithm "cascades" down a series of features and uses a weak classifier to separate non-faces (with high probability) from maybe-faces. The result is an enriched dataset, or in terms of an image, a smaller search region that is more likely to contain a face. This idea of reducing the feature space so that the minority class is no longer quite so rare directly informs this research.

In addition, there are a variety of other methods used in rare event detection problems. Burez and Van den Poel [33] studied how to better handle class imbalance

in churn prediction. Customer churn is when contractual customers or subscribers leave a company. They found that weighted random forests, as a cost-sensitive learner, outperformed the standard random forests algorithm. Similarly, Golub et al [34] also used a series of weighted votes based on correlation to approach the problem of identifying new cancer classes (class discovery) and assigning tumors to known classes (class prediction), using human acute leukemias as a test case.

Others studies used more probabilistic approaches to detecting rare events. Lawrence, Hong, and Cherrier [35] developed two classes of models to predict cabin level no-show rates on airline flights. Airlines routinely overbook flights based on the expectation that some fraction of booked passengers will not show for each flight. This work showed that a new ensemble probabilistic model for predicting rare no-shows increases revenue gain over the conventional model. In contrast, Pednault et al [36] modified a tree-based learning technique to model insurance risks. They adjusted the construction of branch splits to overcome selection biases that arise because of the imbalance that is present in the data.

The literature demonstrates a common theme for detecting rare events: adaptation and flexibility. It is often helpful when dealing with a class imbalance to modify a technique to fit the unique nature of the data. Some studies apply weighted or un-weighted voting. Others adjust how the learning algorithm fundamentally selects the predicted class, or use a sophisticated re-sampling method to balance the data. The best method varies – an approach that may work for cancer classification may not work for a different rare event detection problem.

3.2.3 Performance Measures

For both the baseline and joint methodology, we test the models on real world, out-of-sample data. This allows us to evaluate model performance using a set of metrics. Although accuracy is generally a good indication of success, when 95% of the data belongs to a single class, we can achieve 95% accuracy by simply putting everything into that class! We propose several additional criterion to better measure how the model performs specifically with respect to negative social response.

Training and Test Set Selection

To effectively evaluate the model's ability to predict NSR, we ran out-of-sample tests using previously unseen test data. Using 5-fold cross validation, we randomly split the dataset into five equal parts. Then we use four-fifths of the data to train the model and the remaining one-fifth to test.

We also ensure that the training data is balanced. A balanced dataset has an equal distribution of classes (an equal number of no response and NSR events). When the data is imbalanced, the classification algorithm learns to give preferential treatment to the majority class, to the detriment of the minority. To offset this issue, balancing is achieved through random resampling of the training data, with a bias towards a uniform class. That is, we randomly resample to *reduce* the number of No Response events, and *increase* the number of outbreaks resulting in NSR. The test set is left unbalanced to represent the real world distribution of target outcome values.

Test Set Evaluation

To evaluate model performance, we use the following metrics to better measure how the model performs with respect to the rare class. They will also help show how the joint methodology improves upon the baseline individual supervised learning algorithms.

Table 3.1: Generic confusion matrix

	Predicted None	Predicted NSR
No Response	*true negative*	*false positive*
NSR	*false negative*	*true positive*

1. Accuracy $= \dfrac{\text{number of } \textit{true positives} + \text{number of } \textit{true negatives}}{\text{total number of instances}}$

2. Sensitivity $= \dfrac{\text{number of } \textit{true positives}}{\text{number of } \textit{true positives} + \text{number of } \textit{false negatives}}$

3. Precision $= \dfrac{\text{number of } \textit{true positives}}{\text{number of } \textit{true positives} + \text{number of } \textit{false positives}}$

Although accuracy is often a good overall measure of model performance, it does not contain all the necessary information. For example, consider the following confusion matrix where NSR events are associated with about 5% of the data:

Table 3.2: Example confusion matrix used to illustrate problem of poor precision and poor sensitivity, but good accuracy

	Predicted None	Predicted NSR
No Response	850	200
NSR	20	40

This yields 80% accuracy, but there are five times as many false positives as true positives, which is unacceptable for a biosurveillance detection problem with limited resources in the real world. We must consider that decreasing the rate of false negatives often increases the rate of false positives, and vice versa. When trying to detect a rare event like NSR, *precise sensitivity* is more important than pure accuracy.

3.2.4 Supervised Learning

Both the baseline and joint methodologies use the Random Forest algorithm and Logistic Regression to predict NSR. They are very different methods that are widely applicable and easy to implement.

Random Forests

The Random Forest algorithm is a popular form of supervised learning because of consistent good performance and lowered variance. The algorithm builds a large collection of de-correlated trees, and then averages them [37]. Each tree is a random classification tree that recursively partitions a dataset into smaller and smaller groups using a random subset of the available features, so that each leaf of the tree represents similar values based on the target label. The random selection of features reduces the correlation between trees, thus reducing the overall variance of the model.

Figure 3-8 shows a simple example of a random classification tree using the features "country" and "disease". If an outbreak occurs in Afghanistan, and the disease is

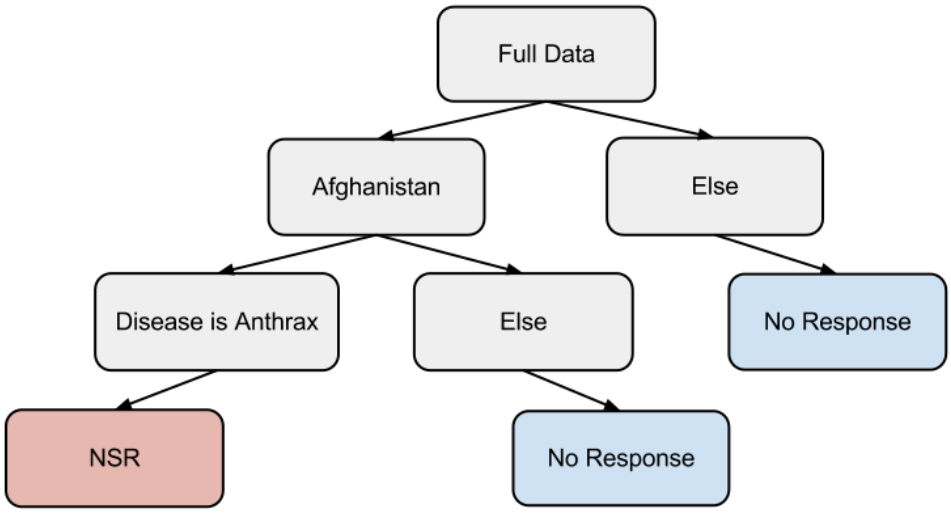


Figure 3-8: Example of random classification tree structure

Anthrax, the algorithm predicts NSR. Otherwise, the algorithm predicts no response. Random Forests builds hundreds of trees like this one using different features, and then averages the results.

For each outbreak in the test set, the many classification trees built by the algorithm outputs a predicted class (in our case, NSR or No Response). Each of these outputs represents one committee vote, where the majority of the votes determines the overall predicted class [37]. Although an individual tree might yield poor accuracy, the combination of many trees often provides very stable and accurate results. To implement Random Forests on the biosurveillance data, we must choose the number of trees to create, and the number of random features to consider.

Logistic Regression

Another type of supervised learning, Logistic Regression, models the posterior probabilities of a binary target label using a linear function of the features [38]. Specifically, the log-odds of a NSR outcome is fit to the features using linear regression. There are no explicit parameters that require tuning.

Consider the tree example used in Figure 3-8 above. If we build a linear regression model using the features "country" and "disease", we obtain a continuous output

variable z. If we then take the logistic function of z, $f(z) = \frac{1}{1+e^{-z}}$, we obtain the probability (between 0 and 1) of the outbreak causing a negative social response. Anything below 0.5 is classified as no response, and anything above is classified as NSR.

3.2.5 Joint Methodology

The *Data-driven Predictive Model* uses supervised learning to analytically predict outbreaks that result in negative social response. Although supervised learning is a powerful classification tool, sometimes its effectiveness is obscured by other factors. Noisiness of the data, rarity of the target label, and the limitations of individual learning algorithms lead us to employ a multi-step, joint methodology. This is a sequential process of adding interaction features to the data, enriching the data, and applying a voting strategy.

Interaction Features

The data is noisy because there is a disparity between what might really be happening within an outbreak, and what is reported by the media/internet. This noisiness causes a large variance in the performance of a learning algorithm, so to help address this issue, we build interaction features. As shown in Figure 3-9, interaction features create combinations of feature values. In this thesis, we created two-variable interactions for only the most relevant attributes, approximately doubling the number of available features. For example, we combined region and an undisclosed binary variable, yielding a new categorical interaction variable with values like Africa.0, Africa.1, NorthAmerica.0, and so on.

Although supervised learning algorithms can detect these interactions, sometimes it is beneficial to highlight them. This increases the importance of interaction features to the model and decreases the importance of the original features. In doing so, we reduce the variance of the model and boost the predictability of NSR.

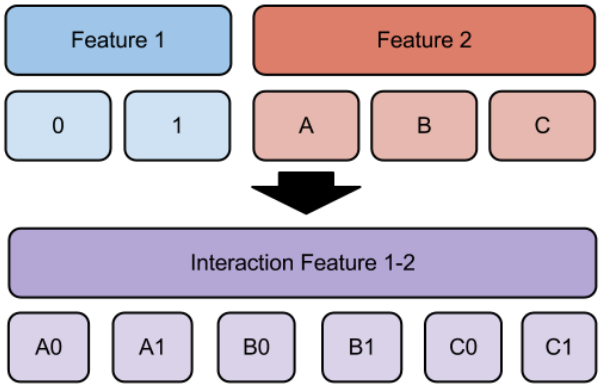


Figure 3-9: Visualization of Interaction Features. Here we combine a binary attribute and a categorical attribute to create a new categorical interaction feature.

Data Enrichment

As previously noted, less than 5% of the data s associated with NSR, and class rarity is a serious problem for any supervised learning algorithm. Because the classifier is burdened with an overabundance of events with no social response, it has trouble distinguishing between relevant and irrelevant outbreaks. For example, all outbreaks of Dengue Fever in Peru generated no response. Further examination of the data shows that there are many such feature values where no response occurs. To take advantage of this fact, we apply a decision tree to identify feature values that are not associated with NSR and classify these outbreaks as No Response. Figure 3-10 illustrates the data enrichment algorithm.

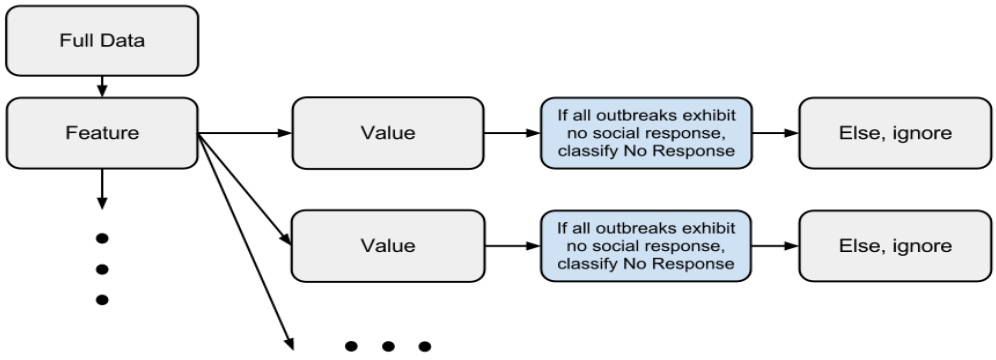


Figure 3-10: Overview of Enrichment Decision Tree. It uses feature values to immediately classify events that are never NSR as No Response.

The result of the above process is a reduction of the feature space, yielding enriched data. Figure 3-11 is a visual representation of this result. By eliminating some events without social response from consideration, NSR is less rare. This allows the algorithm to learn using more relevant data.

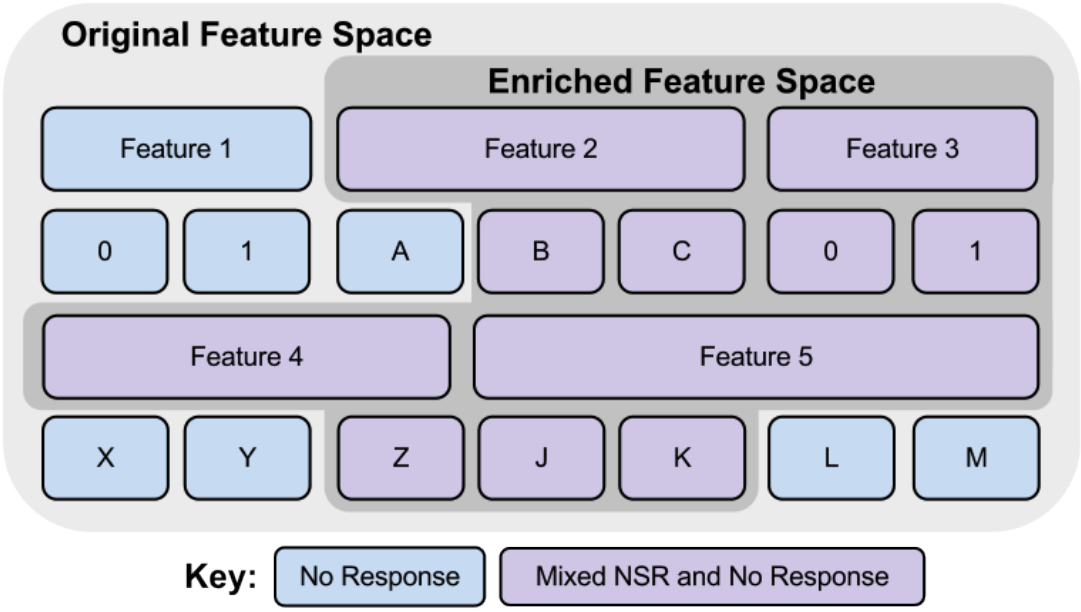


Figure 3-11: Visualization of Enriched Feature Space. The enrichment process yields a dataset where NSR is less rare.

Voting Strategy

We can now apply the supervised learning algorithms to the enriched data. Using a type of ensemble learning, we combine two models to achieve better performance than the individual models could achieve otherwise. On the biosurveillance data, the Random Forest algorithm achieves high precision but low specificity, while Logistic Regression achieves high specificity but extremely low precision. An unweighted voting strategy allows us to combine the high precision of Random Forests with the high specificity of Logistic Regression, by giving each model a single vote. Figure 3-12 shows the voting strategy as a decision tree. We chose this strategy based on the nature of the data and the performance of the individual supervised learning models.

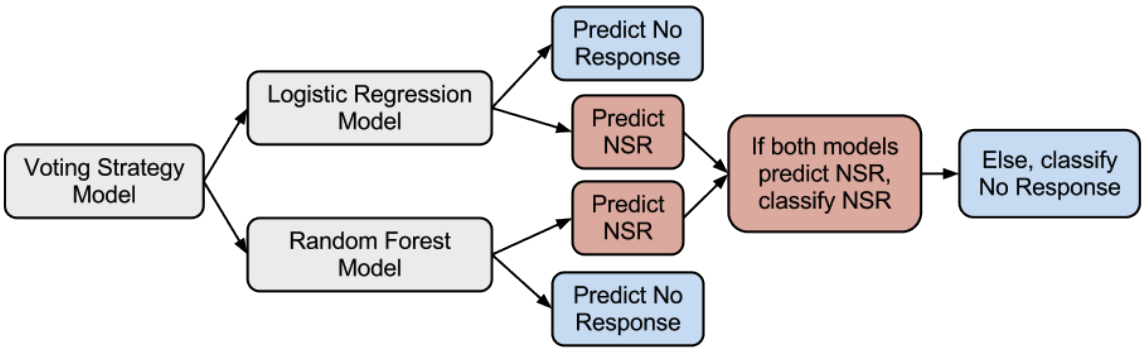


Figure 3-12: Overview of Voting Strategy Decision Tree. It gives the predicted outcome of both models equal weight, and only classifies an event as NSR if *both* models agree.

The *Data-driven Predictive Model* uses the joint methodology to improve the prediction of outbreaks that result in NSR. This sequential process of adding interaction features to the data, enriching the data, and applying a voting strategy helps reduce false positives while maintaining rare class accuracy. The first two steps are primarily concerned with better identification of outbreaks with no response, but only goes so far towards improving the performance of the learning algorithms. The voting strategy is the glue that holds the whole model together; it combines the positive effects of Logistic Regression and Random Forests, thus yielding better sensitivity and precision in predicting negative social response to the spread of an infectious disease.

Chapter 4

Results

In this chapter, we describe the implementation and analysis of the *Epidemic Social Response Model* and the *Data-Driven Predictive Model*. Both models utilize the biosurveillance data governing all outbreaks of Dengue Fever. We chose to narrow the analysis by disease to improve the interpretability of the results and for ease of implementation. We will demonstrate how an agent-based modeling approach, combined with a predictive framework, can help decision-makers better understand and anticipate negative social response (NSR) to the spread of an infectious disease.

The social network analysis in Section 4.1 simulates a network of individuals connected through social ties and analyzes how the interactions of agents spread infection and negative social response. To validate the model, we test it on real world examples from the data outlined in Chapter 2: outbreaks of Dengue Fever in City X and City Y. Recall that City X is an example of a typical outbreak that resulted in no notable social response, while City Y is an example of an extreme negative social response to an outbreak. The model simulation shows that the complex behavior of an outbreak can be successfully explained by a series of simple agent interactions on a network. The results provide some insight into the mechanism that drives the spread of social response to an outbreak.

The data mining approach in Section 4.2 analytically predicts outbreaks that result in negative social response using supervised learning. Because NSR events are rare and detecting them leads to a high rate of false positives, the model applies a

joint methodology to improve performance. This consists of additional interaction features, data enrichment, and a voting strategy. The results will show how each of these components build on each other to yield reasonably accurate predictions of negative social response to the spread of an infectious disease.

4.1 *Epidemic Social Response Model*: Results

To validate the performance of the *Epidemic Social Response Model*, we performed multiple Monte Carlo Simulations. Using the example outbreaks provided in Chapter 2, the interaction probabilities were adjusted so that the behavior of the model matches the behavior of real world data. The simulations will show that mathematical and computational modeling represents an effective tool for exploring the impact of an outbreak on the social response of a population. The process is outlined in Figure 4-1.

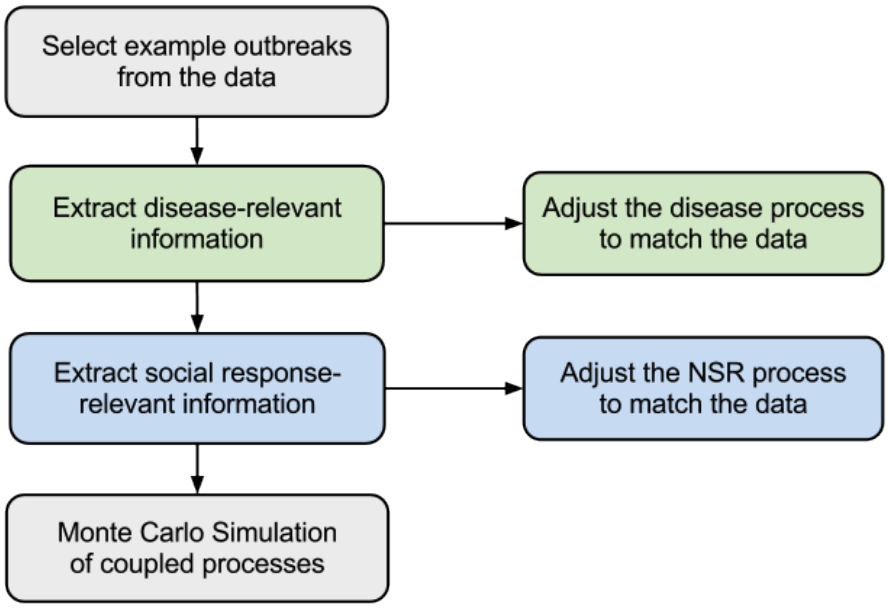


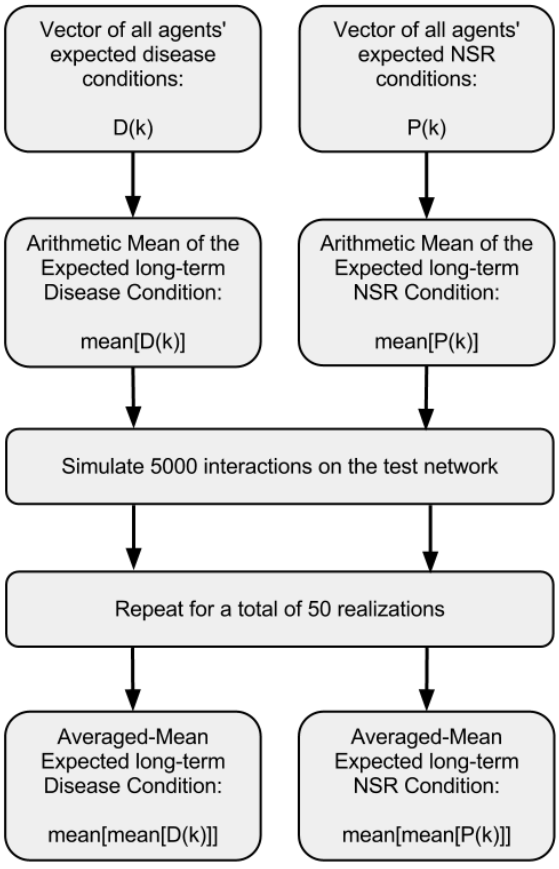
Figure 4-1: Overview of *Epidemic Social Response Model* adjustment, where relevant data is used to verify the model's performance.

4.1.1 Monte Carlo Method

We implement a Monte Carlo Simulation in MATLAB, in order to analyze the condition dynamics of the model on the test network. A Monte Carlo Simulation is an algorithm that relies on replicating the probabilistic behavior of the system using computers [39]. The algorithm simulates the model dynamics at each interaction, employing random sampling to select an agent pair and probabilistically adjust their conditions. This process is repeated over a specified number of total interactions.

4.1.2 Model Initialization and Adjustment

Recall that the model consists of both the NSR process on top of the disease process. The condition of any agent is his NSR level (between 0 and 10) and disease class (susceptible, infected or recovered).



We simulated 5000 interactions on the test network for each example outbreak, producing plots of the arithmetic mean of the expected long-term condition at each interaction, k. We repeat this process for a total of 50 individual realizations, where each realization yields an arithmetic mean. We then average to produce the mean (over all all realizations) of the mean (over all agents) of the expected long-term conditions for each interaction, k. For simplicity, we refer to these calculations as the *averaged-mean* expected long-term condition.

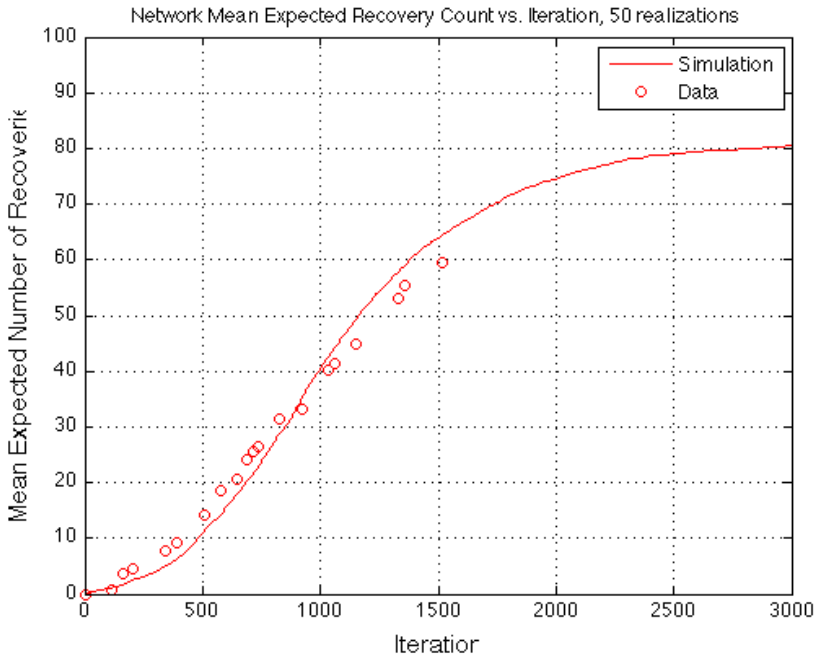
Figure 4-2: Visualization of how Monte Carlo simulation outputs are obtained

For each example outbreak, we have data tracking the progression of the disease, up to the point where the infection rate begins to taper off. This closely follows the overall number of recoveries in the model, data first displayed in Figures 2-6 and 2-7. By appropriately scaling the real world data, we can adjust the probability of infection and the probability of recovery to match the model. Figure 4-3 (City X) and Figure 4-4 (City Y) on the following pages show the data fit to the averaged-mean plot, as well as the large variation in the model. Although both of the averaged-mean plots look very similar despite the differences in the outbreak, this is due to the scalability of this model. An outbreak with 1,000 infections or 10,000 infections can be scaled-down and represented on the same size network. This also makes City X and City Y directly comparable – we will see later how two very similar outbreaks on the same social network can generate wildly different negative social response.

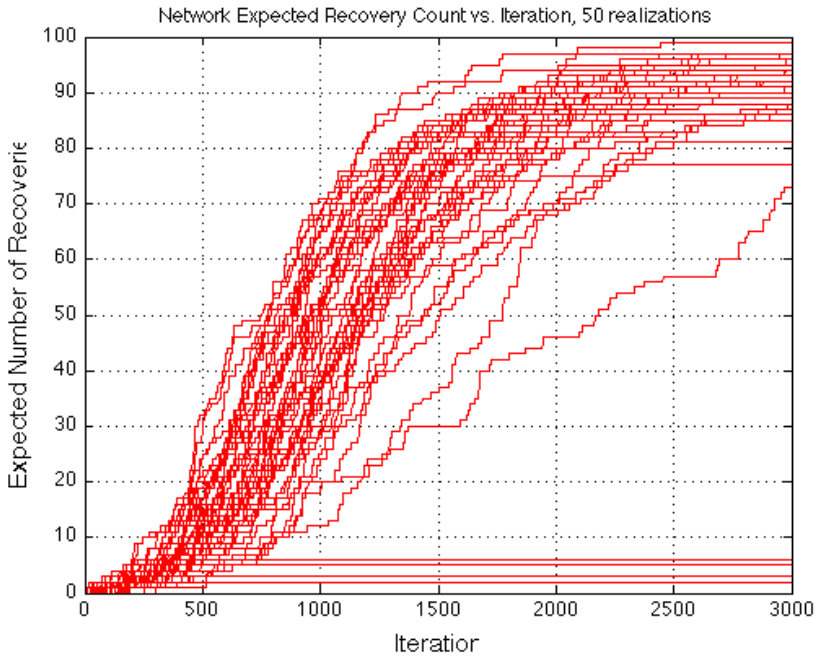
Although the data does not perfectly fit the averaged recovery curve, it is well within the margin of error. The disease process can proceed in a variety of ways. Sometimes, it dies out before spreading through the population. Other times, the infections spread much more slowly, so we see a less exponential and more linear representation of the total number of recoveries. These simulations are important for two reasons. First, they show that our modification of the SIR model (that uses the probabilistic framework of the spread of misinformation model [24]) accurately represents the recoveries in an infectious disease outbreak. Second, the individual simulations of the disease progression illustrate the inherent volatility of any outbreak.

The adjustment of the disease process and the recovery portion of the model is a fairly straightforward operation. The detailed nature of the outbreak updates in the biosurveillance data allows us to extract very specific reports of the event's total number of cases. However, there is no such data about the spread of NSR – we only know roughly when the negative response peaks, and that severe or widespread NSR should at a minimum exceed the average condition (above level 5). From this we can adjust the NSR process in a more rudimentary way.

Figure 4-3: **City X** Mean and Averaged-mean expected long-term Number of Recoveries, versus Interaction Number.

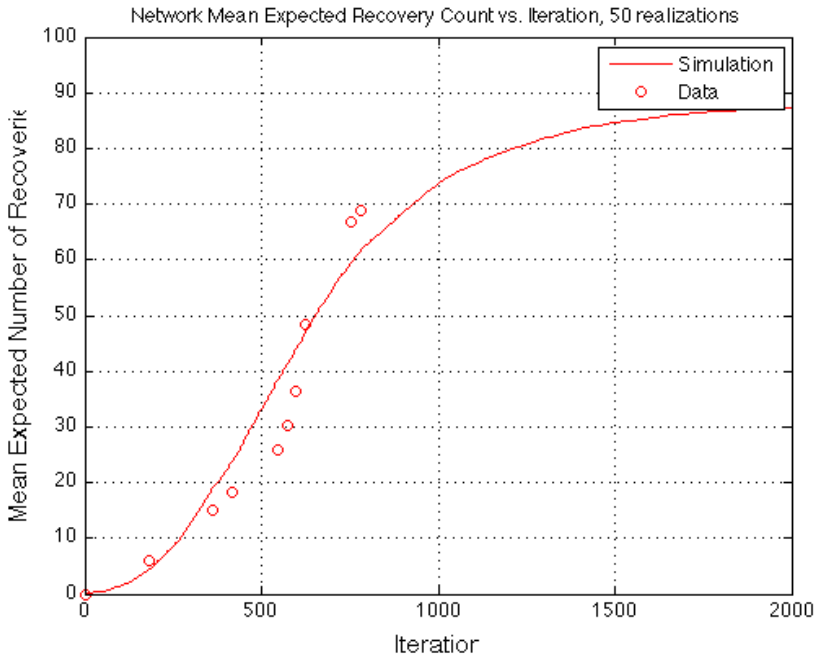


(a) The averaged-mean plot follows a typical recovery curve for an SIR model. The data is extracted from the total number of reported cases for the outbreak in question, scaled to fit the timeframe, and the disease parameters are adjusted to match.

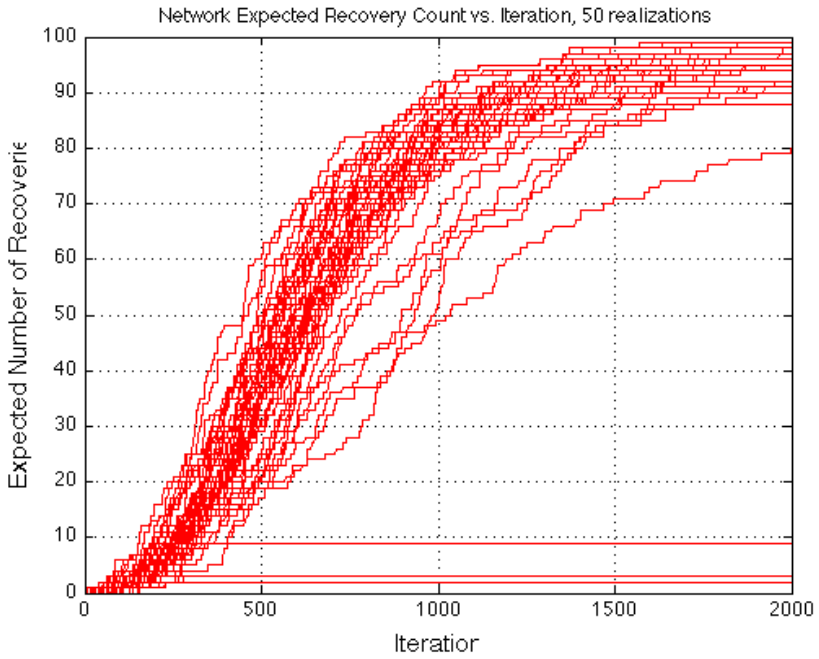


(b) The plot of the individual means shows a wide variation about the averaged curve above, indicating that the less-than-perfect fit of the data is well within the margin of error.

Figure 4-4: **City Y** Mean and Averaged-mean expected long-term Number of Recoveries, versus Interaction Number



(a) The averaged-mean plot follows a typical recovery curve for an SIR model. The data is extracted from the total number of reported cases for the outbreak in question, scaled to fit the timeframe, and the disease parameters are adjusted to match.



(b) The plot of the individual means shows a wide variation about the averaged curve above, indicating that the less-than-perfect fit of the data is well within the margin of error.

4.1.3 Simulation Results

After fitting the data to the model, we produce plots for the disease and NSR processes. It is important to note that simulating the model can only provide us with general insights; analytic results are left for future research into the subject.

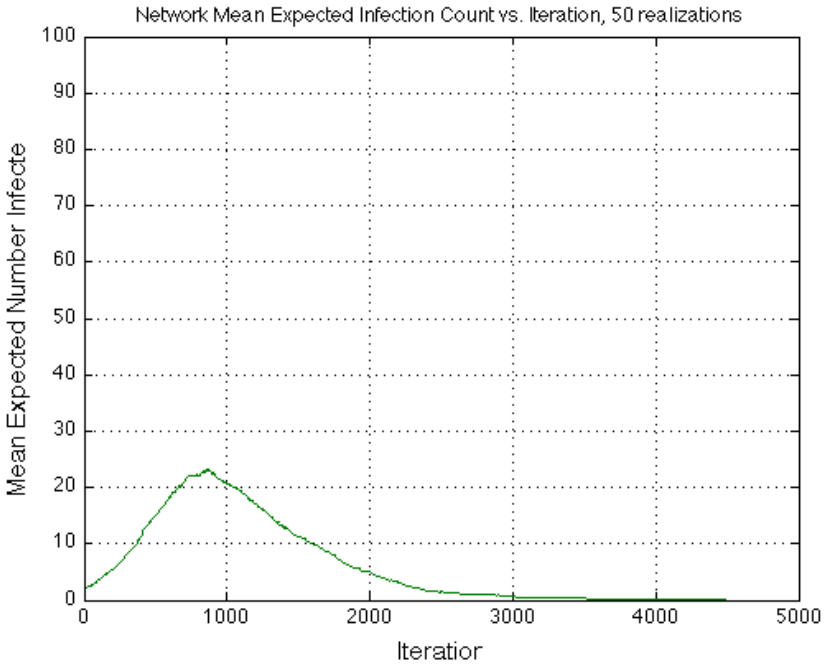
The Average Behavior of the Model

Figure 4-5 shows the outbreak of Dengue Fever in City X. At the peak of the disease spread, less than one-quarter of the population is currently infected. It experiences a steady rise and decline in the number of new infections, consistent with what the data indicates happens over the course of the outbreak. The associated negative social response is negligible.

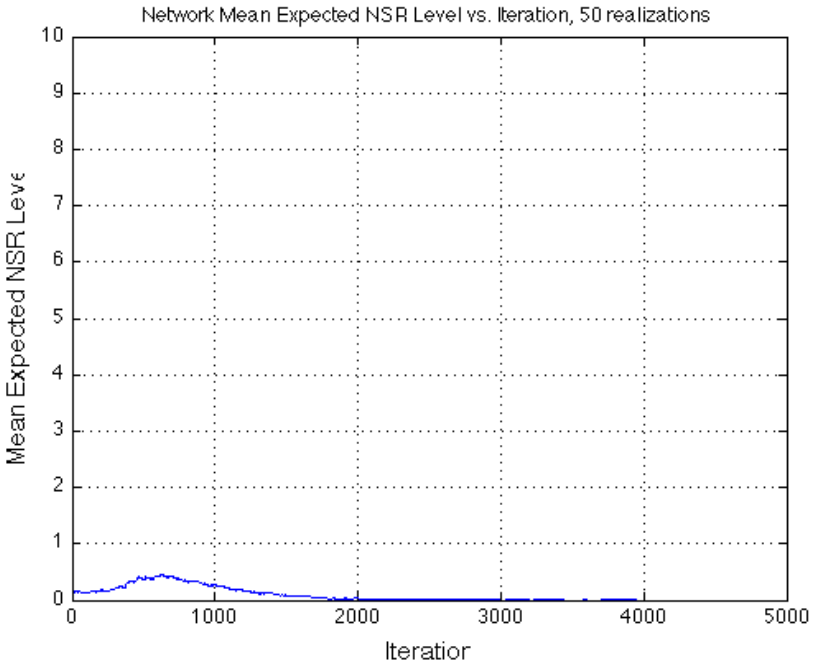
In contrast Figure 4-6 shows the outbreak of Dengue Fever in City Y. The peak of the disease spread occurs earlier with a much steeper climb in new infections. It also requires fewer iterations for the outbreak to run its course. Again, this behavior is consistent with what the data indicates happens over time. City Y's outbreak is both more sudden and shorter in duration than that of City X.

The behavior displayed in these simulations is achieved only by adjusting the interaction probabilities of the processes. To achieve a shorter, more sever outbreak, we increase the probability of infection (ν) and the probability of recovery (κ), and reduce the difference between them. To achieve a severe negative social response, we primarily increase the probability of forcefully spreading NSR (α_{HI}). To achieve a negligible social response, we set all parameters to zero but for decay (δ), which is set to 1. This means that NSR *never* spreads amongst the population – the only time it is introduced into the model is through the coupling effect. For instance, in the City X outbreak, people develop a negative social response when they become infected, but they only calm down over time and never spread it to others. On the other hand, in the City Y outbreak, panicked agents are very "contagious" to others and only calm down slowly, so the NSR spreads quickly and persists.

Figure 4-5: **City X** Averaged-mean expected long-term Number of Infections and NSR Level, versus Interaction Number

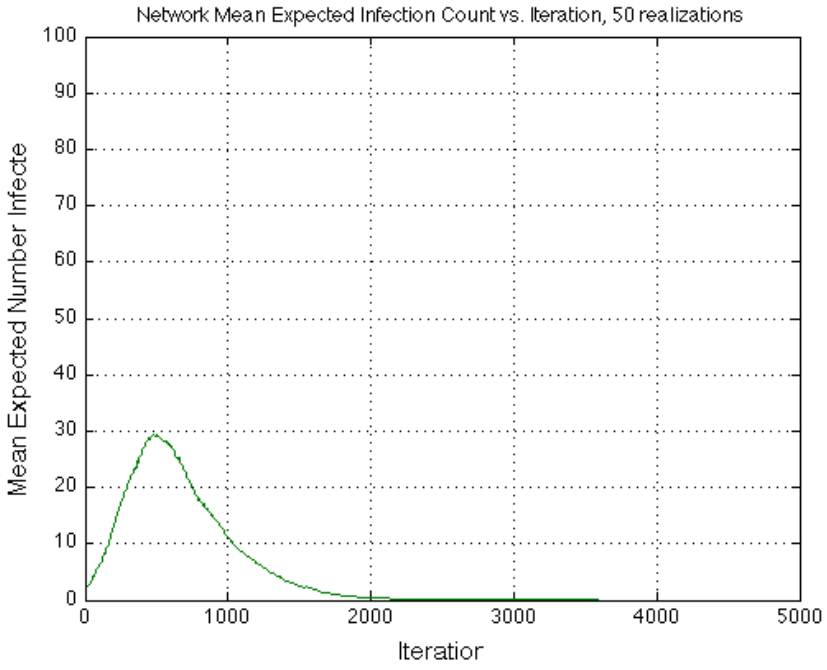


(a) A mild spread of an endemic infection

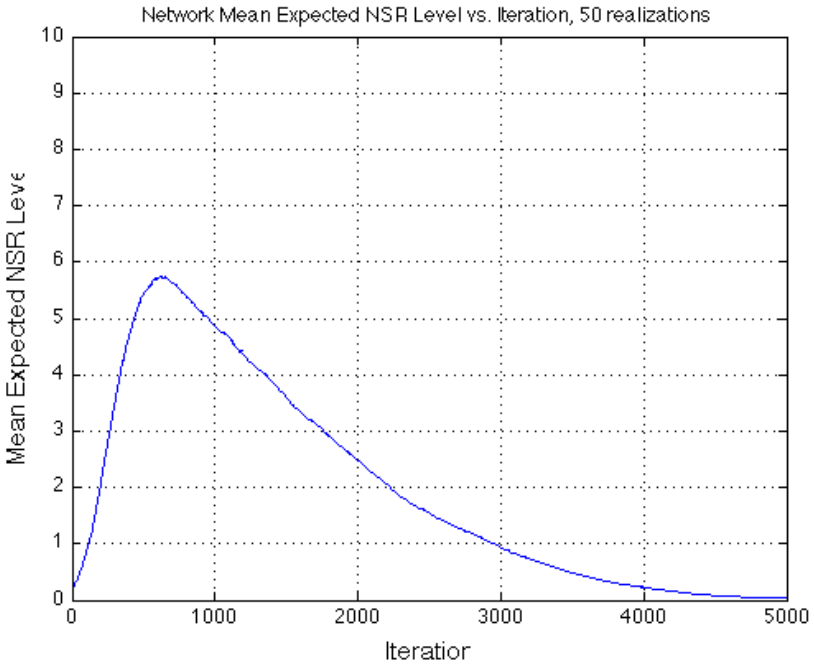


(b) The negligible social response to the disease

Figure 4-6: **City Y** Averaged-mean expected long-term Number of Infections and NSR Level, versus Interaction Number

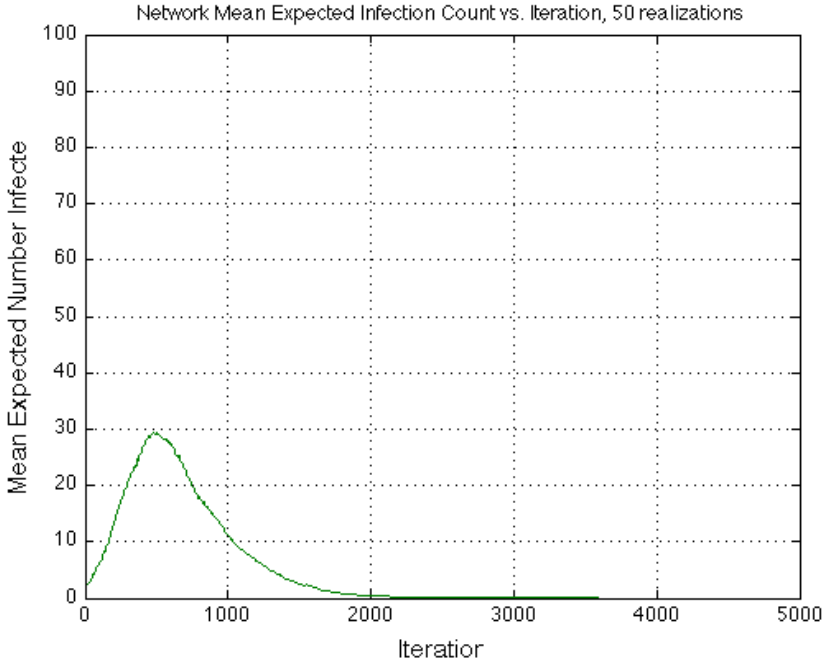


(a) A quicker and more severe spread of an infection

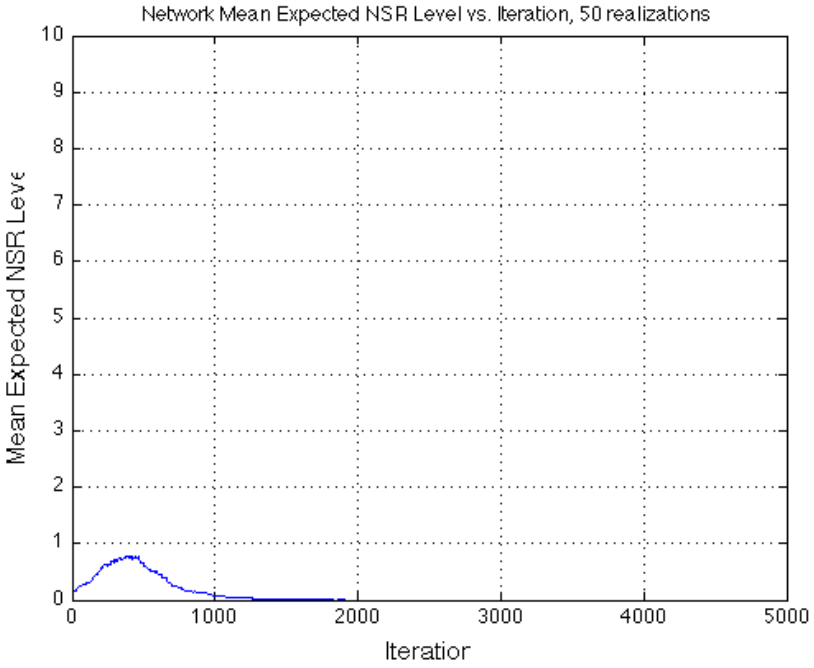


(b) The severe negative social response to the disease

Figure 4-7: **Sensitivity** Even with the more severe disease spread of City Y, NSR can be very mild. It is a volatile and nearly independent process from the outbreak.



(a) City Y's disease spread



(b) Mild negative social response to the disease, achieved using the same parameters for the NSR process as City X. Compare to Figure 4-5b – there is only a tiny increase in NSR due to the more severe disease spread.

Figure 4-7 further illustrates how the interaction probabilities affect the model. What happens when an outbreak displays the more severe disease spread of City Y, but *does not* exhibit a negative social response? The simulations demonstrate that although the disease does slightly impact the maximum NSR level, the negative response is almost completely independent of the disease spread. That is, by adjusting the NSR interaction probabilities, we can achieve a mild social response regardless of the type of outbreak. The coupling mechanism serves to tie the two processes together, but allows enough separation that we can model a wide variety of situations.

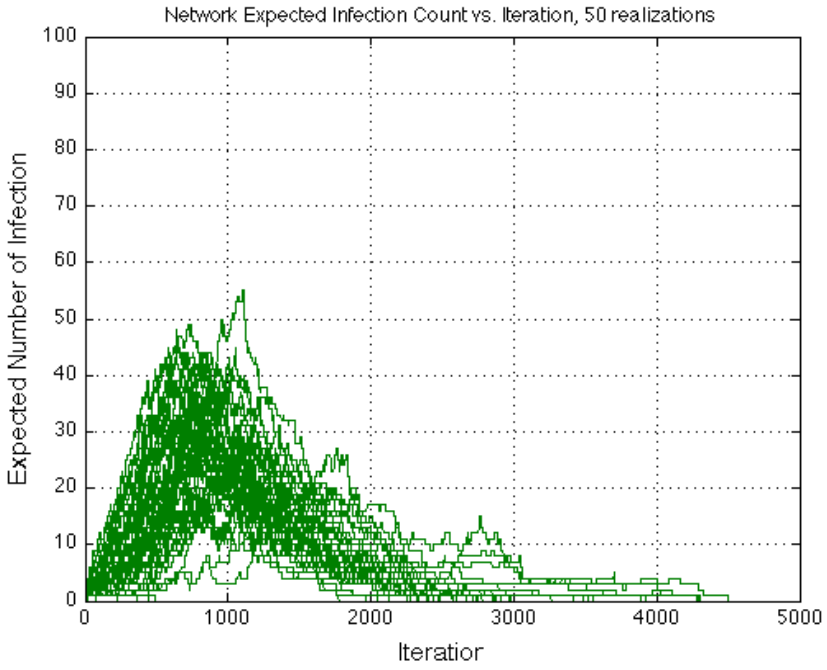
The implication of these parameter adjustments is the most important result of the social network analysis. The model is based on repeated agent interaction over a network, relying on a set of probabilistic rules. This simplified representation of individual behavior, a hallmark of agent-based modeling, can effectively capture the collective behavior of an outbreak and whether the affected population shows strain associated with the disease.

The Variation of the Model

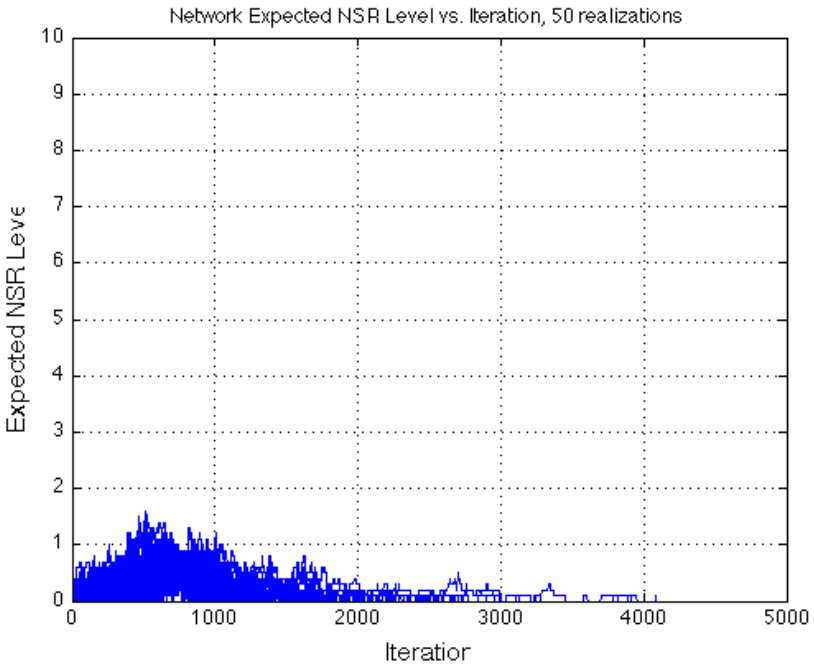
Because the model is probabilistic and highly random, there is a large variation in behavior. Sometimes, the disease never reaches critical mass and dies out before spreading through the population. In such an event, even though NSR is coupled to the disease, once infected an agent can disseminate negative social response independent of what happens to the outbreak itself – NSR may die out, or it may continue to spread. Alternatively, sometimes despite widespread infections in the population, the NSR is only maintained at a low level. Any number of situations can occur, depending on which agents interact and how they exchange conditions.

Figure 4-8 shows the variation in City X's outbreak of Dengue Fever. Notice the outliers in the disease process – some simulations last longer, others shorter; some peak above 50% of the population while others never reach critical mass. The NSR process is consistently negligible but still shows incredible variation.

Figure 4-8: **City X** Mean expected long-term Number of Infections and NSR Level, versus Interaction Number

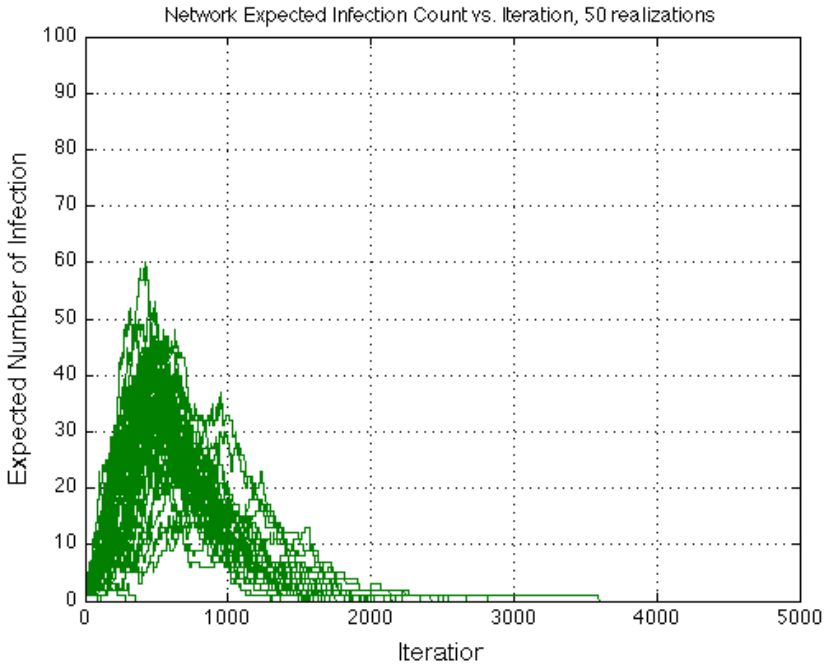


(a) City X Infection spread for 50 realizations of the model

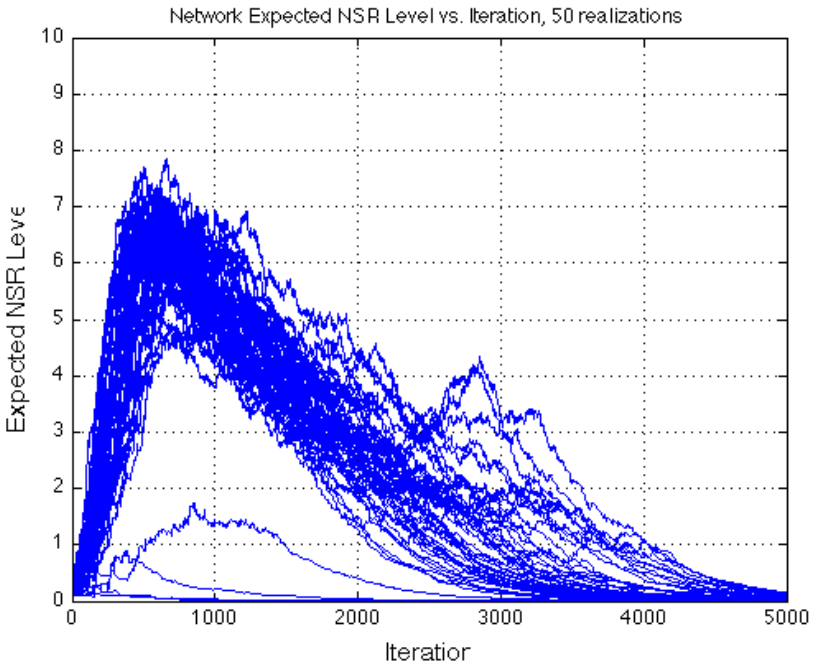


(b) City X Negative Social Response spread for 50 realizations of the model

Figure 4-9: **City Y** Mean expected long-term Number of Infections and NSR Level, versus Interaction Number



(a) City Y Infection spread for 50 realizations of the model

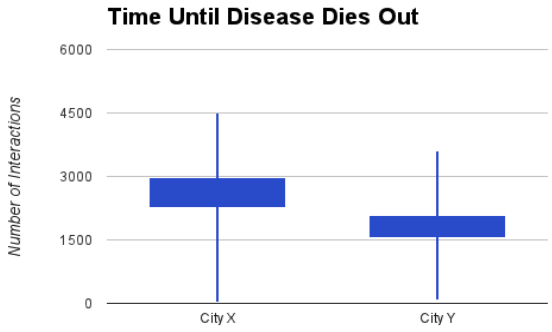


(b) City Y Negative Social Response spread for 50 realizations of the model

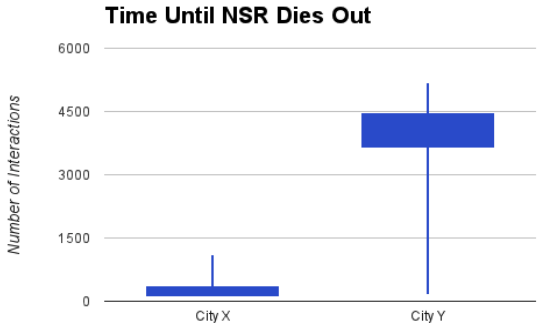
Similarly, Figure 4-9 describes the variation in City Y's outbreak. Here the infection rate almost always sharply increases due to the high probability of infection. However, there are some interesting outliers in the NSR process. In cases where the disease *did* die out quickly, the NSR *did not*! There are also cases where a late spike in anxiety showed up long after the disease was on the decline.

The variation in the model is further indication that the *Epidemic Social Response Model* can accurately portray real world outbreaks. In addition to plotting the simulations, we also calculated some basic summary statistics to better display the differences in the example outbreaks: the time until the disease process dies out, the time until the NSR process dies out, the maximum number of infections, and the maximum negative response level.

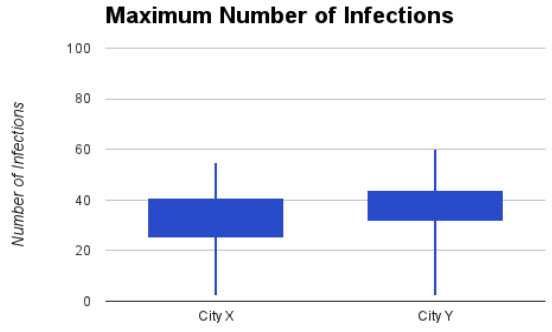
Figure 4-10: Box-and-whisker plots of statistics gathered from 50 realizations of the model. Despite little difference in the disease metrics, the NSR metrics show more variation. In addition to having greater levels of NSR, City Y also shows greater extremes about the mean.



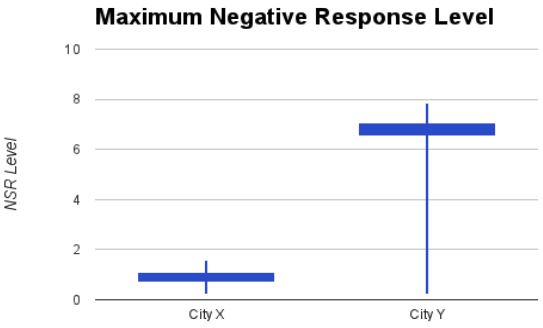
(a) Outbreak X has a slightly longer duration than Y, though similar variance.



(b) Outbreak Y shows persistent NSR with a large variance, unlike Outbreak X.



(c) Outbreak X has a slightly milder infection rate than Y, though similar variance.



(d) Outbreak Y shows severe NSR with a large variance, unlike Outbreak X

Figure 4-10 is a box-and-whiskers plot that shows the middle 50%, the min, and the max of the simulation realizations for each of these statistics. They confirm what we already know, that City Y had a slightly shorter and more severe outbreak than City X, and that the negative social response was extremely high on average (but frequently quite low in the extremes).

Together, these insights verify what the experts describe as the primary forces that drive negative social response. When faced with an unusual outbreak, the lack of knowledge builds and multiplies until the population begins to show serious strain. If the pathogen is highly infectious or has an increased fatality rate, the NSR effect is amplified. Future research in this area will examine the analytic effect of an outbreak on social response, and use the model as a tool to help decision-makers save resources and limit negative social response to the spread of an infectious disease.

4.2 *Data-Driven Predictive Model*: Results

4.2.1 Data Selection

To increase the interpretability of the results, and for ease of implementation, the model considers only outbreaks of Dengue Fever (DF). There are 1,096 initial outbreaks of Dengue Fever recorded in the biosurveillance data with no social response; 46 are associated with indicators of negative social response, for an incidence rate of 4%. The worldwide distribution of the infectious disease, by country, is shown in Figure 4-11; dengue is generally concentrated in rainy, equatorial regions of the world.

Of all the diseases in the data set, Dengue Fever was chosen for two reasons: *(a)* it had the second highest number of outbreaks, and *(b)* the NSR rate was close to the overall incidence rate.

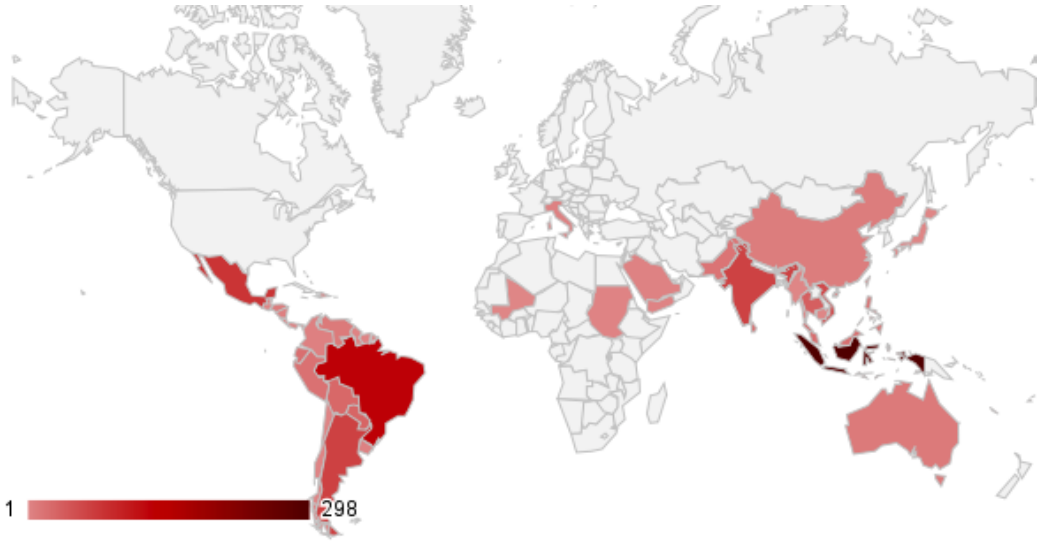


Figure 4-11: Heatmap – Number of Dengue Fever outbreaks by country

4.2.2 Overview of Predictive Performance

The results are broken down into two main sections: baseline methodology and joint methodology. For the baseline methodology, we perform two experiments. The first experiment trains and tests the Random Forests and Logistic Regression models on balanced data, which helps determine whether negative social response is indeed predictable, given a world where NSR and No Response outbreaks are equally likely. The second experiment trains on balanced data, but test the classifiers on skewed real world data where NSR events are very rare. The joint methodology outlined in Section 3.2 follows the structure of the second experiment, but creates interaction features, enriches the data, and applies a voting strategy to improve upon the baseline performance.

Figure 4-12 shows an overview of the prediction results (for the first two cases, we show the average performance of Random Forests and Logistic Regression models). All accuracy, sensitivity, and precision metrics are color coded throughout for viewing convenience; red indicates poor performance, yellow acceptable, light green good, and dark green indicates excellent performance.

Method	Accuracy	Sensitivity	Precision
Baseline - balanced	69.6%	73.9%	68.2%
Baseline - unbalanced	83.8%	45.7%	29.7%
Joint	95.8%	56.5%	48.1%

Figure 4-12: Overview of prediction methodology results. In an ideal situation with balanced data, we have a high rate of detecting NSR. However, with real-world rare event data, performance plummets. By applying the joint methodology, we show significant improvement.

We can draw several basic conclusions from the results overview. The balanced baseline experiment demonstrates that when NSR events are not obscured by an overwhelming number of events with no response, NSR is predictable. However, when NSR is very rare, as it is in reality, it is much more difficult to detect. Both Random Forests and Logistic Regression models poorly predict the target label – on average, these individual learning algorithms never exceed 50% precision or sensitivity. On the other hand, if we apply the joint methodology, we see significant improvement in all three performance metrics. More details on the baseline methodology and joint methodology are discussed in the following sections.

4.2.3 Performance of Baseline Methodology

Experiment 1 – Training and Testing Balanced Data

A balanced set is used to train the classifiers, as it has been shown to improve predictive performance on the test set. This experiment yields 46 NSR outbreaks of Dengue Fever, and 46 no response outbreaks. We then build a Random Forests model and a Logistic Regression model, as described in Section 3.2.4. The results are shown in Figure 4-13. Random Forests in particular yields a very strong, balanced confusion matrix, and correctly predicts 73.9% of events, compared to 65.2% correctly predicted by Logistic Regression.

This experiment is an important exercise for several reasons. To non-experts NSR seems like a random event. By looking at the data alone (not the detailed reports), it is unclear why one outbreak is associated with a negative social response

Figure 4-13: Detailed Performance for Baseline Methodology, applied to balanced data.

	Predicted None	Predicted NSR
No Response	27	19
NSR	13	33

(a) Logistic Regression

	Predicted None	Predicted NSR
No Response	33	13
NSR	11	35

(b) Random Forests

Method	Accuracy	Sensitivity	Precision
Logistic Regression	65.2%	71.7%	63.5%
Random Forests	73.9%	76.1%	72.9%

(c) Performance Metrics show NSR is more predictable than expected

and another is not. Building a model on balanced data demonstrates that given the right information, an algorithm can achieve what we alone cannot – over 70% accuracy for all performance metrics, demonstrating that negative social response is algorithmically predictable.

Experiment 2 – Testing on Unbalanced Data

We now move on to the full, unbalanced dataset. Using the methodology outlined in Section 3.2.3, we build a balanced training set, and apply a Random Forests model and a Logistic Regression model to the withheld data in the unbalanced test set where NSR is rare. The results are shown in Figure 4-14.

Overall accuracy has increased: Logistic Regression went from 65.2% to 71.5% accuracy, while Random Forests jumped from 73.9% to 96.1% accuracy. However, this is due to the swell in the number of no response events that are much easier to predict. It is more revealing how the other performance metrics have changed:

- Although Logistic Regression maintains high sensitivity, the model predicts over ten times as many *false positives* as *true positives*.
- The opposite is true for Random Forests, which maintains a high precision at the cost of many *false negatives*.

Figure 4-14: Detailed Performance for Baseline Methodology, applied to unbalanced data.

	Predicted None	Predicted NSR
No Response	791	305
NSR	20	26

(a) Logistic Regression

	Predicted None	Predicted NSR
No Response	1081	15
NSR	30	16

(b) Random Forests

Method	Accuracy	Sensitivity	Precision
Logistic Regression	71.5%	56.5%	7.9%
Random Forests	96.1%	34.8%	51.6%

(c) Performance Metrics highlight the flaws in the models: either many false positives, or many false negatives. The complementary nature of these results will be exploited by the joint methodology.

The duality in the results is the driving force behind the joint methodology. Step one, creating interaction features, and step two, enriching the feature space, are both designed to improve the individual model performance of Random Forests and Logistic Regression. We can then take further advantage of the strengths of each model by combining their results in a third step using a voting strategy.

4.2.4 Performance of Joint Methodology

As described in Section 3.2.5, the joint methodology is a multi-step approach to improving the baseline performance. Figure 4-15 demonstrates the step-by-step improvement in performance.

For Logistic Regression, creating interaction features reduced the false positives from 305 to 212 and the false negatives from 20 to 13. Enriching the data then further improved the specificity. For Random Forests, the methodology helped decrease the false negatives from 30 to 24 to 15, without drastically increasing the false positive rate. This incremental increase in performance is what led us to apply the sequential joint methodology, and the complementary nature of the results implies that a voting strategy is the next logical choice.

Figure 4-15: Interim Performance for Joint Methodology, showing the step-by-step reduction in false negatives and false positives.

	Predicted None	Predicted NSR
No Response	884	212
NSR	13	33

(a) Logistic Regression with Interaction Features

	Predicted None	Predicted NSR
No Response	1064	32
NSR	24	22

(b) Random Forests with Interaction Features

	Predicted None	Predicted NSR
No Response	876	220
NSR	10	36

(c) Logistic Regression with Interaction Features and Enrichment

	Predicted None	Predicted NSR
No Response	1040	56
NSR	15	31

(d) Random Forests with Interaction Features and Enrichment

Figure 4-16: Final detailed Performance for Joint Methodology, showing significant improvement of the a single supervised learning model.

	Predicted None	Predicted NSR
No Response	1068	28
NSR	20	26

(a) Confusion Matrix

Method	Accuracy	Sensitivity	Precision
Joint	95.8%	56.5%	48.1%

(b) Performance Metrics highlight that the voting strategy virtually eliminates the problem of high false negatives and false positives

Figure 4-16 illustrates the last effect of the voting strategy. Although this study attempted a variety of voting strategies for multiple models, "2 out of 2" remained the best: an event is only ultimately classified NSR when *both* models predict negative social response. The final result preserves the sensitivity of a Logistic Regression model, with the precision of a Random Forests model. We can precisely predict over half of all NSR events.

These results indicate the best performance the *Data-driven Predictive Model* can achieve on unseen test data. More typical results show some variation. Although the overall accuracy remains extremely stable, with an average of 96% and a standard deviation of 0.3%, the other metrics vary moderately. Sensitivity and Precision typically range from 43% to 55%. The ability of the classifier to identify NSR events shows the most volatility, which is expected due to the rarity of these events and the opacity of the underlying reasons why negative social response occurs in the first place.

Although typical results are slightly worse, the optimal results presented above are important for several reasons. Out of approximately 1100 cases of Dengue Fever, we can reduce the "search space" for NSR outbreaks by a factor of 20. This research illustrates that algorithms can provide actionable intelligence, by identifying a small group of outbreaks where one in two are associated with negative social response. Furthermore, this can be done in a matter of minutes, while an expert would require days to search each outbreak individually and determine which will yield a social response. Another result of this research is its applicability to post-analysis. Using an algorithmic approach, we can identify what the data is lacking and other ways it can be improved to take advantage of automated detection capabilities. The predictive model presented in this research will enable decision makers to identify problematic outbreaks and intervene as necessary. Most importantly, it provides the foundation for expanded biosurveillance and improved detection of negative social response to the spread of an infectious disease.

4.3 Summary of Results

In summary, we observe that the results of the social network analysis and the data mining approach form a comprehensive view of negative social response. The simulation results validate the behavior of the agent-based model outlined in Chapter 3, and prove that a simple system of rules applied to agents on a network can effectively capture the collective behavior of a complex epidemic. It help us to

understand the mechanism which drives social response to an outbreak – it is like any other idea and spreads much like a disease itself, from person to person. The predictive model forms a macro view of the problem, with the primary result that the joint methodology of adding interaction features, enriching the data, and applying a voting strategy improves NSR predictions over the baseline methodology. We also note that although negative social response can be predicted with high accuracy in an ideal world, the rarity of NSR events in the real world causes a very high rate of false positives and false negatives. The success of the predictive model is its ability to produce good results despite the rare event detection problem.

Chapter 5

Conclusions & Future Research

5.1 Conclusions

This work provides a foundation for an expanded biosurveillance capability. Traditional biosurveillance is primarily concerned with anticipating and detecting bio-events in order to limit the direct impact of the disease, such as loss of life. We propose an additional focus on the *population strain* associated with the the spread of an infectious disease. Negative social response (NSR) can have far reaching impacts on the social and economic stability of a population. To minimize this impact, it is important to characterize NSR and understand how and when it occurs in the real world.

In Chapter 2, we gave an in-depth description of the data. In particular, we described the rare event detection problem, that less than 5% of the data are associated with negative social response. Furthermore, not only is NSR rare, but it shows incredible variation by both region and disease. To better understand how NSR presents in a community, we also selected two example outbreaks of Dengue Fever that are referenced throughout this thesis. City X was a mild, endemic event that did not inspire a negative response, while City Y was a much more severe outbreak, with contradictory reports and widespread NSR. With a better understanding of the data, we developed two models in Chapter 3 to describe social response to the spread of an infectious disease. The models utilized real world biosurveillance data in different

ways, to analytically predict negative response to an outbreak and to understand how that response spreads within the outbreak.

The *Epidemic Social Response Model* successfully modeled a population as a social network and analyzed how the interactions of agents spread infection and negative social response. The formulation of the model is important for several reasons. We believe it to be the first of its kind – although many other studies exist that individually model disease or social influence, the interplay of the two processes is a fresh approach that provides unique insight into the driving mechanism behind NSR. It spreads much like a disease itself– the more "contagious" NSR is, the more severe the response. When people are faced with an uncommon situation, like a new disease or an unusually high fatality rate, their anxiety spreads throughout the network, making the process takes on a life of its own. Together, both layers of the model – the NSR process on top of the disease process – help give a more complete picture of an epidemic. Future research will be able to take the model a step further, and study not only how the disease influences NSR, but how social response in turn affects the disease.

The model is also extremely flexible, allowing us to represent a variety of real world situations. The key simulation results validated the model formulation and demonstrated that the complex behavior of an outbreak can be explained by a system of basic probabilistic rules. As the simulation of Dengue Fever outbreaks in City X and City Y showed, the model can accurately capture real world bio-events. Furthermore, just a few parameters is enough to model the data. By adjusting the probability of infection and the probability of recovery, we can control the severity and duration of the disease process, matching almost any typical outbreak of infectious disease. By primarily adjusting the probability of one agent forcefully panicking another, we can create either mild or severe negative social responses. In combination, the interaction probabilities are all that is needed to model real world data. The *Epidemic Social Response Model* is a fresh, interesting, and flexible analysis of social response to the spread of disease.

The *Data-driven Predictive Model* predicts what types of outbreak in general cause a social response. We developed a joint methodology to improve performance over a more conventional supervised learning algorithm, by creating interaction features, enriching the data, and applying a voting strategy. Interaction features and data enrichment helped the learning algorithms to better identify outbreaks with no response, by increasing model specificity and creating a dataset where NSR is less rare. The voting strategy combined the positive characteristics of Logistic Regression and Random Forests. The joint methodology as a whole improved the detection of rare NSR events through the reduction of false positives and false negatives, yielding reasonably accurate predictions of negative social response to the spread of an infectious disease. Out of over 1000 Dengue Fever events, we precisely predicted over half of the rare NSR occurrences. Although seemingly random to a non-expert, these results indicate that NSR is algorithmically predictable. An expanded biosurveillance system could be successfully implemented to identify and mitigate these adverse events in the future.

In addition to better detection of negative social response, the predictive model is also important for post-analysis. It helps guide future data collection; for example, variables with low predictive power can indicate that the feature or feature values need to be redefined to better capture the real world behavior of an outbreak. Using an algorithmic approach, we can identify what the data is lacking and other ways it can be improved to take advantage of automated detection capabilities. The predictive model presented in this research will enable decision makers to better characterize and predict infectious disease events, in order to mitigate the adverse effects.

5.2 Future Work

Future research should include several additions and modifications to both the social network model and the predictive model.

Updates to the *Epidemic Social Response Model*:

- Provide analytic solutions. It would be useful to predict the maximum NSR level achieved by the model, the time until the process dies out, or how quickly the disease or NSR diffuses through the population, given a set of fixed interaction probabilities.

- Apply the simulations to other real world examples and expand the sensitivity analysis of the parameters.

- Create separate contact networks for the disease process and the NSR process, and compare the performance of the model under a variety of network topologies. Increase the size of the test networks and introduce different classes of agents (with varying levels of influence on their neighbors).

- Examine alternative disease processes, such as SEIR, MSEIR or SIS. Depending on the disease being studied, some epidemiology models are more accurate than others. For instance, SIS is often better for endemic infections.

- Examine the effect of policy actions on negative social response and on the infections. For example, how do media blasts, quarantines, or vaccinations impact the model?

Updates to the *Data-driven Predictive Model*:

- Expand the analysis beyond Dengue Fever to other diseases and regions.

- The data also includes a free text description of each event. Future research should extract any relevant additional information for predicting NSR.

- Analyze how negative social response spreads between outbreaks. Use the data updates to predict whether an outbreak will develop NSR in the future, or whether future unrelated outbreaks will develop a response due to a past NSR event.

While the results in this work currently exist at the fringes of biosurveillance, the field is rapidly growing and diversifying. The increasing availability of data and growing need for stronger detection capabilities are driving ever greater expansion. This study provides a framework for the future examination of many different disease consequences, in addition to social response. We are confident that this work is merely a foundation of much more research to come.

Bibliography

[1] M. Wagner, A. Moore, and R. Aryel. *Handbook of Biosurveillance*. Elsevier Academic Press, 2006.

[2] Biosurveillance Coordination Unit. National biosurveillance strategy for human health. Technical Report 2, Centers for Disease Control and Prevention, February 2010.

[3] Centers for Disease Control and Prevention. Bioterrorism overview. `http://www.bt.cdc.gov`, 13 April 2012.

[4] Gayathri Vaidyanathan. Better biosurveillance could halt disease spread. `http://www.nature.com`, 29 June 2011.

[5] James Simpson. German e-coli traced to bean sprouts - bacteria contains plague dna. `http://www.examiner.com`, 11 June 2011.

[6] Deborah Sontag. In haiti, global failures on a cholera epidemic. `http://www.nytimes.com`, 31 March 2012.

[7] Merco Press. Bolivia dengue cases could reach 50,000 by end of march. `http://en.mercopress.com`, 3 March 2009.

[8] Channel News Asia. One more victim dies in geylang food poisoning incident. `http://www.channelnewsasia.com`, 8 April 2009.

[9] U.S. Department of Health and Human Services. Assessment of the 2009 influenza a (h1n1) pandemic on selected countries in the southern hemisphere:

Argentina, australia, chile, new zealand and uruguay. http://www.flu.gov, 26 August 2009.

[10] The Sydney Morning Herald. Mob stones chilean bus amid flue fears. http://news.smh.com.au, 23 May 2009.

[11] Matthew Jackson. *Social and Economic Networks*. Princeton University Press, Princeton, NJ, 2008.

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[13] Ascel Bio. Home. http://www.ascelbio.com/, April 2012.

[14] Centers for Disease Control and Prevention. Provisional surveillance summary of the west nile virus epidemic. *Morbidity and Mortality Weekly Report*, 20 December 2002.

[15] Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *PNAS*, 99(3):7280–7287, May 2002.

[16] G.H. Lewes. *Problems of Life and Mind (First Series)*. Trubner, London, 1875.

[17] M. E. J. Newman. The structure and function of complex networks. *SIAM*, 45(2):167–256, 2003.

[18] Sandro Meloni, Nicola Perra, Alex Arenas, Sergio Gomez, Yamir Moreno, and Alessandro Vespignani. Modeling human mobility responses to the large-scale spreading of infectious diseases. *Games and Economic Behavior*, 1(62), August 2011.

[19] W. Kermack and A. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London.*, 115(772):700–721, 1927.

[20] S. Ma and Y. Xia. *Mathematical Understanding of Infectious Disease Dynamics*. Lecture notes series. World Scientific, 2009.

[21] Elihu Katz and Paul Lazarsfeld. *Personal influence: the Part Played by People in the Flow of Mass Communications*. Free Press, Glencoe, IL, 1955.

[22] Robert P. Abelson. Mathematical models of the distribution of attitudes under controversy. In *Contributions to Mathematical Psychology*, chapter 6, pages 141–160. Holt, Rinehart, and Winston, Inc., New York, NY, 1964.

[23] Morris H. DeGroot. Reaching a consensus. *Journal of American Statistical Association*, 69(345):118–121, March 1974.

[24] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. Spread of misinformation in social networks. *Games and Economic Behavior*, 70(2):194–227, November 2010.

[25] Bernard Waxman. Routing of multipoint connections. *Selected Areas in Communication*, 6(9):1617–1622, December 1988.

[26] Dimitri Bertsekas and John Tsitsiklis. *Introduction to Probability*. Athena Scientific, Belmont, MA, second edition, 2008.

[27] Torgny Lindvall. *Lectures on the Coupling Method*. Courier Dover Publications, 1992.

[28] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[29] G.A. Carpenter, S. Grossberg, N. Markuzon, J.H. Reynolds, and D.B Rosen. Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, 3:698–712, 1992.

[30] Joseph Downs, Robert Harrison, and Simon Cross. Evaluating a neural network decision-support tool for the diagnosis of breast cancer. In Pedro Barahona, Mario Stefanelli, and Jeremy Wyatt, editors, *Artificial Intelligence in Medicine*,

volume 934 of *Lecture Notes in Computer Science*, pages 239–250. Springer, Berlin/Heidelberg, 1995.

[31] Michael Kearns. Thoughts on hypothesis boosting. Machine Learning class project, December 1988.

[32] Jianxin Wu, James Rehg, and Matthew Mullin. Learning a rare event detection cascade by direct feature selection. Technical Report GIT-GVU-03-16, Georgia Institute of Technology, 2003.

[33] J. Burez and D. Van den Poel. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36:4626–4636, 2009.

[34] et al. T. R. Golub. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(531), 1999.

[35] R. Lawrence, Se June Hong, and Jacques Cherrier. Passenger-based predictive modeling of airline no-show rates. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 397–406, New York, NY, 2003. ACM.

[36] Edwin P. D. Pednault, Barry K. Rosen, and Chidanand Apte. Handling imbalanced data sets in insurance risk modeling. Technical Report WS-00-05, AAAI Technical Report, 2000.

[37] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, chapter 15. Springer, 2009.

[38] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, chapter 4. Springer, 2009.

[39] Reuven Rubinstein and Dirk Kroese. *Simulation and the Monte Carlo Method*. Wiley, second edition, 2008.