Thesis

# Identification and Characterization of Super-spreaders from Voluminous Epidemiology Data

Submitted by

Harshil Shah

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2016

Master's Committee:

Advisor: Shrideep Pallickara
Co-Advisor: Sangmi Pallickara

F. Jay Breidt

## Abstract

### Identification and Characterization of Super-spreaders from Voluminous Epidemiology Data

Planning for large-scale epidemiological outbreaks often involves executing compute-intensive disease spread simulations. To capture the probabilities of various outcomes, these simulations are executed several times over a collection of representative input scenarios, producing voluminous data. The resulting datasets contain valuable insights, including sequences of events such as super-spreading events that lead to extreme outbreaks. However, discovering and leveraging such information is also computationally expensive. In this thesis, we propose a distributed approach for analyzing voluminous epidemiology data to locate and classify the super-spreaders in a disease network. Our methodology constructs analytical models using features extracted from the epidemiology data. The analytical models are amenable to interpretation and disease planners can use them to inform identification of super-spreaders that have a disproportionate effect on epidemiological outcomes, enabling effective allocation of limited resources such as vaccinations and field personnel.

Table of Contents

# List of Tables

# List of Figures

CHAPTER 1

# Introduction

According to the Food and Agricultural Organization (FAO), there are currently more than 1.5 billion cattle, 1.1 billion sheep, 0.97 billion pigs and goats, and nearly 20 billion chickens in the global livestock industry. This industry employs at least 1.3 billion people around the world, accounting for nearly 18% of the world population [1]. In 2002, the infectious disease Ebola had killed around 5000 critically endangered western gorillas at the Lossi Gorilla Sanctuary located in northwestern Republic of the Congo [2]. Further, nearly three quarters of the rural human population and one third of the urban population depend on livestock directly or indirectly for food, income, transportation or any other services [3] [4]. An issue that accompanies the large human dependency on livestock is that humans are critically susceptible to zoonoses —infectious diseases of animals that can naturally be transmitted to humans [5].

In 1918, a deadly influenza pandemic infected 500 million humans and wiped out nearly 50 million members of the world population [6]. Effective planning for livestock management and the control of infectious threats to farm animals are extremely pivotal for maintaining an intact ecological system, the global economy and human health. Successful planning and resource allocation during disease outbreaks is best accomplished by identifying premises (a group of animals) that are likely to become super-spreaders (premises that are disproportionally infecting other premises). Prevention is always better than a cure, but alleviating the severity of disease is often the only option left once a disease has started spreading through the population. Effective planning involves the timely use of limited resources to target the super-spreading premises involved in disease outbreaks because super-spreaders are often

highly responsible for the severity of a given disease within a population. One of the recent super-spreading event was Severe Acute Respiratory Syndrome (SARS), which started in China and spread through 37 countries within two weeks.

The degree of human dependency on livestock and the severe consequences of disease incursion have led to significant efforts on the part of the epidemiological modeling community to understand and predict the distribution of disease within an animal premise as well as its transmission within premises [7]. Epidemiological models, often expressed as stochastic discrete event simulations, involve hundreds to thousands of input biological and other parameters, and they tend to be compute-intensive. In this thesis, we generated a simulated disease spread network using the North American Animal Disease Spread Model (NAADSM), which has been vetted by over 300 epidemiologists and veterinarians, and is one of the key tools used by the US Department of Agriculture to plan for disease incursions [8]. NAADSM can be used to model foot and mouth disease (FMD), highly pathogenic avian influenza, swine flu, and pseudo-rabies [9] [10] [11]. NAADSM generates a voluminous disease outbreak dataset by considering multiple input parameters and completing multiple simulation runs.

This thesis contributes to pinpointing the most influential premises in disease spread network as super-spreaders that could contribute disproportionately to disease spread (i.e., once particular premises are infected, the total number of infections and the probability of the diseases becoming endemic are all high). Classifying super-spreading premises can be key factor when developing an effective response plan, and determining specific super-spreading premises helps limited resources (vaccines, field personnel, and bio-surveillance) to be allocated in an effective and targeted fashion. Further, identifying effective individualized features that make a premise a super-spreader can give valuable insights to foreseeing

epidemic effects. In this thesis, our analysis is focused on voluminous data from simulation runs and tracking disease evolution through a population.

## 1.1. Scientific Challenges

The timely identification and characterization of super-spreader premises in voluminous epidemiological data introduces a set of unique challenges:

- **Dataset Size:** An epidemiological state is distributed over a large number of files. Each simulated time step produces an output file containing a variety of simulation data that must be processed to capture the disease spread pattern.
- **Timeliness:** The analysis workflow must execute in parallel across a cluster of computing resources to ensure timely results.
- **Scalability:** The proposed methodology must be scalable with increases in the number of premises and interconnectivity between entities for ensuring the generalizability of the approach.
- **Accuracy and Interpretability:** The analysis must be reasonably accurate and support interpretability by explaining why particular premises are considered the super-spreaders. This is critical for fine-tuning outbreak responses.

## 1.2. Research Questions

Research questions that we explore in this thesis are the following:

RQ1: **How can we characterize the influential premises?**

This involves discovering the epidemic characteristics of influential premises as well as the features that comprise these characteristics, enabling interpretability and herd classification.

RQ2: **How can we support efficient analysis over a voluminous dataset?**

Specifically, analytic workflow should be executed in a distributed fashion to extract information from a voluminous dataset (having more than 3M example scenarios) and provide overall knowledge extraction by considering all possibilities. Further, our methodology must be scalable with increases in the number of premises for analysis.

## 1.3. Overview of Approach

Our epidemiology dataset encompasses multiple representative scenario variants and iterations, which we processed to extract and record millions of infection incidents. This includes tracking the number, source, destination, depth of disease transmission, and so on. In this thesis, our analysis of voluminous epidemiology data was two-fold. First, we empirically classified the super-spreaders from a disease spread network with the use of a custom hierarchical aggregation distributed framework. We then conducted a premise-based exploration of properties that contribute to the super-spreading event with the use of machine learning technology. Specifically, to determine super-spreaders in the disease network, we used the Pareto Principle [12] which is highly applicable to super-spreading events and states that approximately 20% of infected premises are responsible for 80% of causality. Then we modeled the relationship between premises based on features extracted from the simulated dataset to classify the super-spreaders using a stochastic gradient descent algorithm and tuned it according to minimizing misclassification [13]. We evaluated and validated the results against the influential premises found via a network analysis algorithm, PageRank [14] [15].

## 1.4. Thesis Contribution

Thesis contributions include:

- A general distributed hierarchical aggregation system workflow for empirical analytics involving millions of data files comprising different disease scenarios.
- A model for highly accurate classification of super-spreaders using the Support Vector Machine (SVM) applied with Stochastic Gradient Descent (SGD) methodology to maximize likelihood. This model can improve resource allocation by identifying pivotal premises.
- Support for interpretability of the analysis by identifying key features that characterize super-spreading premises.

## 1.5. Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 outlines the simulation and dataset used in this thesis, followed by related methodology in Chapter 3. Chapter 4 describes the followed methodology. Methodology includes empirical identification of super-spreaders using a custom distributed hierarchical aggregation framework, and classification of super-spreaders using premise-based properties characterization. Chapter 5 provides a thorough evaluation of our methodology. Finally, conclusions and future research directions are described in Chapter 6.

CHAPTER 2

# Background Information

In this chapter, we describe the tool we used for generating simulated epidemiology datasets: NAADSM. We present the definition of Super-spreaders within the context of epidemiology, and the link-analysis algorithm used for pinpointing influential nodes in the network.

## 2.1. NAADSM

The North American Animal Disease Spread Model (NAADSM) is a spatially explicit, stochastic, state transmission simulation for the spread of highly contagious disease in animals [8]. It was developed with international support to aid strategy development and decision-making for disease attacks. In this model, groups of livestock, called premises, are the basis of simulations. Note that we also use the terms unit and herd to refer to a group of animals.

NAADSM takes several input biological or non-biological parameters into consideration before generating a disease simulation. Disease spread between premises is influenced by production types (i.e. goat, swine), inter-group similarities (shipment rates, infection rates, etc.), relative locations, and geological distances between premises. When a unit is infected, it follows a natural cycle of disease states consisting of: susceptible, latent, sub-clinically infectious, clinically infectious, naturally immune, vaccine immune, and destroyed. This cycle can be interrupted by disease control strategies including quarantine, destruction, vaccination effectiveness and veterinarian visits. Disease spread among premises can happen in any of three ways: direct contact, indirect contact, and airborne spread. Stochastic processes drive all operations in the model, and the processes are based on user-defined distributions and

relational functions. NAADSM input parameters can be of six types: binary values (yes/no), integers, floating point numbers, probabilities, probability density functions, and relational functions.

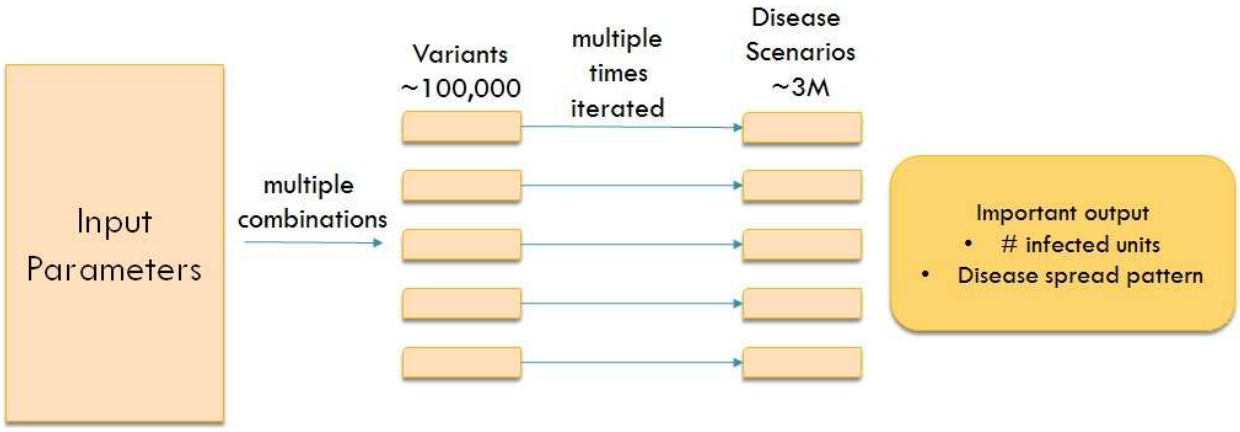



FIGURE 2.1. NAADSM workflow.

The general NAADSM workflow is described in Figure 2.1. Collectively, from different combinations of the input parameters, multiple variants files are generated. Since the simulation is stochastic, each set of input parameter variants is executed several times (32 in this study) to gain statistical confidence in the results. These iterations contribute to the overall representativeness of the output variables probability distributions. Key outputs used during planning include disease duration, number of infected animals, and the disease spread pattern. To reduce the overall execution time of the simulation, NAADSM can be parallelized [16] over a cluster of computing resources in a fault-tolerant fashion [17].

## 2.2. Super-spreading Premises within Disease Spread Networks

In epidemiology, super-spreaders are a phenomenon that is widely observed in disease outbreaks. A super-spreader is an infected unit that spreads a disease disproportionally to other herds [18]. For a given outbreak, there may be more than one super-spreader and the majority of individuals infect multiple secondary contacts. The most recent SARS

outbreak is a considerably notorious example of a super-spreading event (SSE) [19]. Consider the following example of a super-spreading event scenario of how super-spreading premises work.

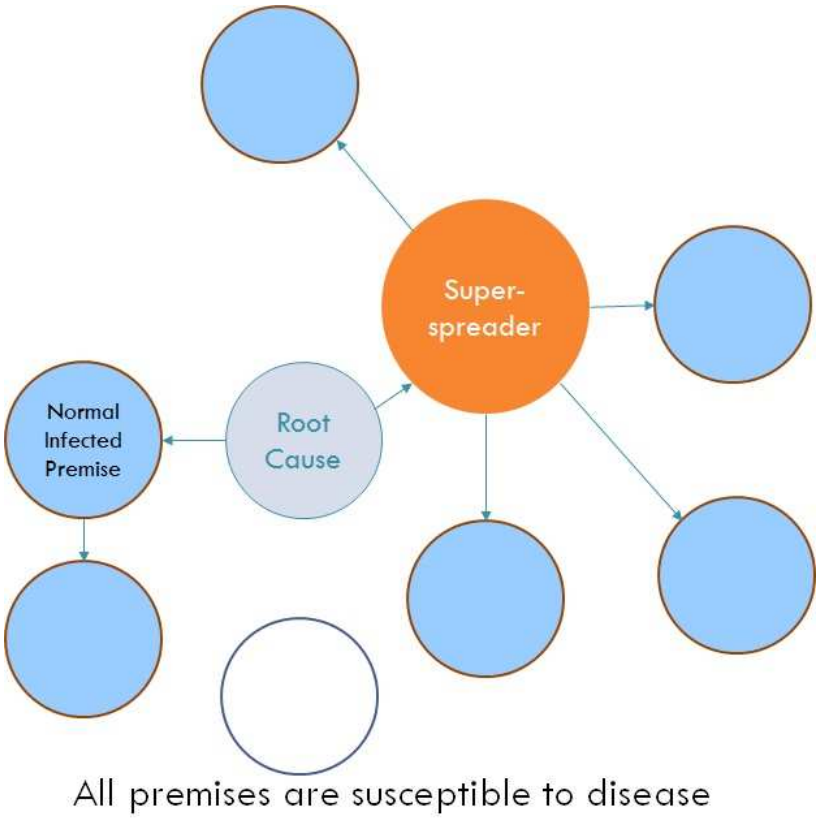



FIGURE 2.2. Example of Super-spreading Event (SSE).

During disease transmission, any unit(s) can be the root cause of a disease that starts travelling, and the disease may reach susceptible premise(s), that due to their intrinsic properties, further infect more secondary premises than the average infected normal premise can infect. In the scenario shown in Figure 2.2, a super-spreading premise infects 4 secondary premises, whereas other normal premises infect none or few.

## 2.3. Subject Dataset

Our subject dataset was derived from a sensitivity analysis that explored the NAADSM parameter space to produce multiple valid combinations of inputs set in Colorado, USA [20] [21]. This process generated 100,000 scenario variants that were executed 32 times for a total of 3.2 million outputs (6.26 TB). The Colorado scenario contains premises from different types of production including swine, goats, sheep, cow-calf and beef. In this particular scenario, a single initial premise was infected, and the disease spread, eventually encompassing tens of thousands of premises. The output of the simulation contains attributes representing the disease status of individual premises and how the infection spreads across premises within the network. These outputs also account for topological characteristics such as connectivity between the premises, proximity, and contact due to movements, but the analysis is bounded with respect to disease transmission from premise to premise (it can be direct, indirect, or airborne).

## 2.4. Influential Premises Identification Using a Link Analysis Algorithm —PageRank

The most influential herds play a vital role in transmitting disease to other susceptible premises disproportionally within a disease network. In these situations, the influence of a given premise depends on the influence of the premise it has infected. In other words, a premise has high influence if it is infecting other highly influential premises. Shah et al. [15] suggested this type of behavioral interaction is efficiently modeled by the link analysis algorithm —PageRank. PageRank was proposed by Page et al. [14] and used by the Google search engine to sort search results by their relevance or importance. In this thesis,

super-spreading premises found via our model are compared and validated against influential premises detected by the PageRank algorithm.

CHAPTER 3

# Related Work

Super-spreaders make disease outbreaks more severe. Analysis of influence in epidemiology has seen considerable study, with much of the work revolving around the various characteristics of infected entities and their impact on disease transmission [22] [23]. Substantial effort has been devoted to identifying hotspots (influential premises) that make diseases result in super-spreading events (SSEs).

Lloyd-Smith et al. define a protocol to identify super-spreaders, which is applicable in understanding SARS outbreaks too [24]. The protocol suggests that the mean number of secondary infections (reproductive capacity $R_0$) from a particular host follows a Poisson distribution and outliers (can be tuned with respect to mean and determined threshold) are often accountable for super-spreading events. However, use of the traditional metric $R_0$ is an inadequate indicator of whether or not an SSE will be triggered because of underestimation of the epidemics potential occurrence even when field observations of mean secondary infections are considerably low [25].

Social Network Analysis (SNA) focuses on human interactions in social networks, but can be applied to epidemiology and in our case for analyzing animal epidemics as well [26] [27] [28] [29]. Fujie et al. focus on intrinsically strong herd infectiousness and social connections [30]. It is always beneficial to include intrinsic strength in such premise-based behavior, but our particular dataset generated using NAADSM simulation, does not reveal premise-based information, so it is irrelevant when we are dealing with an abstract network.

Kitsak et al. proposed a k-shell algorithm for network analysis [31]. The algorithm groups all nodes into $k$-shell values that have $k$ (or less) connections or that are only connected to

other nodes with $k$ (or less) connections only. According to this algorithm, $k$-shell values are assigned in a linear fashion and nodes that reside in the core of $k$-shells (nodes with higher $k$-shells) are considered as the most influential units in the network. The algorithm is time consuming in that it has $O(n^2)$ time complexity. The algorithm cannot be applied directly to a large network having more than 100,000 premises.

Research scientists at the AT&T research lab, Feng et al. [32] modified the $k$-shell algorithm for SNA. Instead of assigning k values linearly, they assigned the values in a logarithmic pattern. Although the algorithm converged faster than the original one, they ended up with a limited list of influential premises. Super-spreaders tend to follow the Pareto Principle, and the target for this thesis is to find nearly 20% of the premises that are super-spreaders, so this approach is unlikely to meet the needs of those who must prevent and manage SSEs.

The PageRank algorithm proposed by Page et al. [14], is famous for its link analysis algorithm, and it is widely used by Google for web search result sorting. The PageRank algorithm gives a weight to each node by requiring higher level internode communication and a greater number of iterations to get converged. [33] We validated our results with the PageRank algorithm, and the results were convincing and interpretable. Weng et al. extended PageRank by considering topical similarity between the users and the link structure between the users [34]. But, although it is dealing with highly abstract graphic structure, such information cannot be used directly.

Cha et al. classify the influential users in Twitter based on three metrics: in-degree, retweets, and mentions. This approach uses Spearman's rank correlation coefficient to compare user influence, and evaluates the behavior of the three metrics for highly influential users [35]. Another approach proposed by Khrabrov et al. [36] uses the daily mentions of

users on Twitter as a basis for calculating different rank metrics such as PageRank, drank, and StarRank to determine influence.

The Hirsch index (or $H - index$) is used in the scientific community for measuring the productivity and impact of a scientist [37]. This algorithm assigns an $H - index$ $i$ to a user, if $i$ of his messages have been retweeted or mentioned at least $i$ times each. Considering a computation intensive premise-based task, the algorithm is not scalable with increases of premises.

CHAPTER 4

# Methodology

Our goal for this study is to identify and classify high-level super-spreading premises in a disease outbreak network. To achieve this goal, we have composed a workflow that comprises two analysis phases. Figure 4.1 represents an abstract view of the approach. First, we classified the premises based on their likelihood to be super-spreaders. Specifically, we performed empirical analysis and found super-spreaders from a voluminous epidemiology dataset. Second, we performed localized classifications to detect herds that have a particularly strong influence on another herd but not necessarily the system as a whole, meaning that, we focus on classifying super-spreaders by studying their epidemic attributes and modeling the relationships between the characteristics. We perform validation and evaluation in chapter 5.

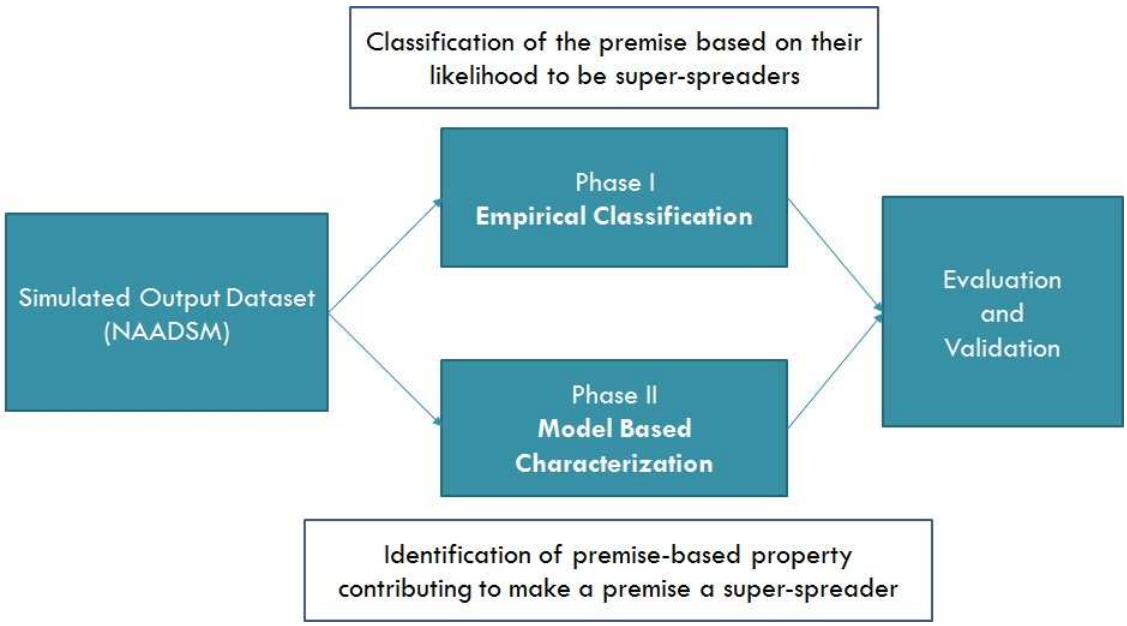



Figure 4.1. Approach summary.

### 4.1. Phase I: Empirical Classification of Super-Spreaders

Super-spreaders tend to follow the Pareto principle [38], also known as the 80-20 rule, where approximately 20% of infected individuals are responsible for 80% of causality [12]. In general, a premise is considered to be a super-spreader if it is responsible for a significantly larger percentage of transmission than a normal infected premise [24]. In this thesis, we used simulated disease scenarios (∼3 million) to classify potential super-spreaders. All premises within the state were considered susceptible to disease. We measured the per-premise infection contribution ($cont_{premiseID}$) to the disease spread network by the influence of each premise on each scenario ($cont_{premiseID-scenarioID}$). Figure 4.2 shows the mathematical workflow for this approach.

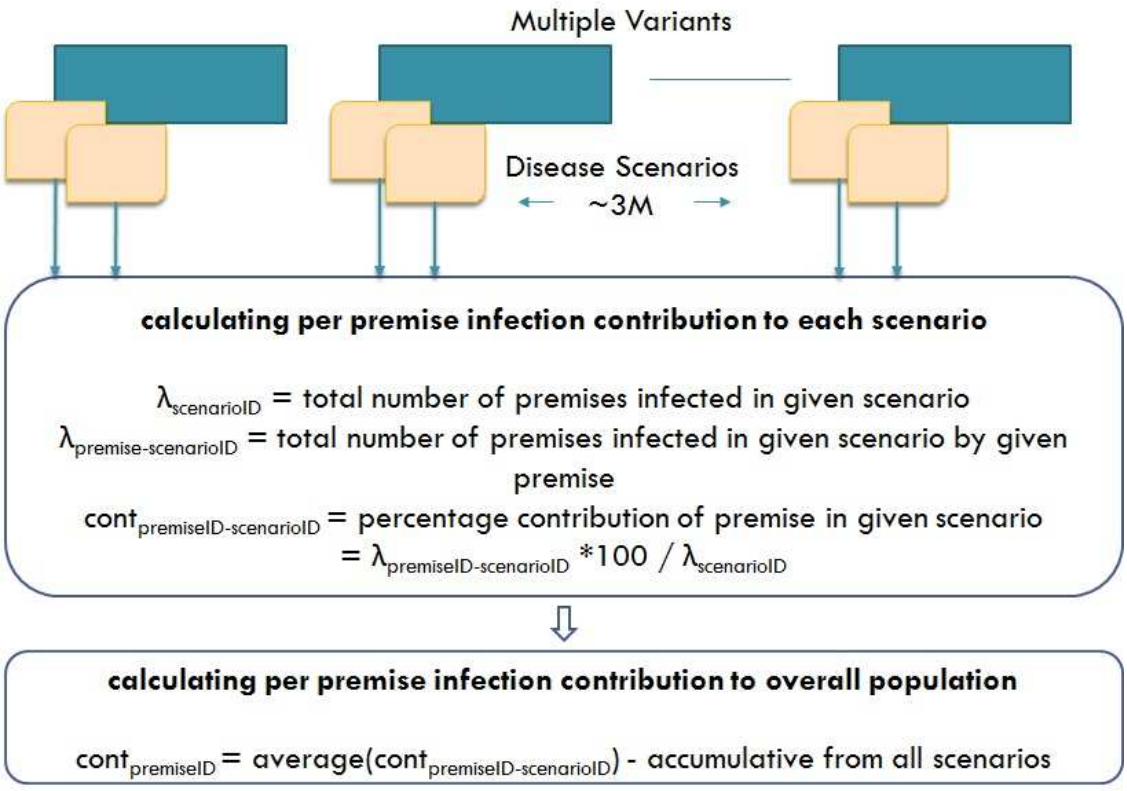



Figure 4.2. Calculating per premise infection contribution to the population ($cont_{premiseID}$).

We applied the 80-20 rule (Pareto Principle) to select the top 20% of premises from a descending-ordered list based on $cont_{premiseID}$ and flagged them as probable super-spreaders. Also, this approach gives every premise a chance to be a super-spreader, except for the seeders in scenarios. The seeders are the premises which start the disease in population. If seeders are considered in counting, then obviously they will show up first in the probable super-spreaders list, so we did not include $cont_{premiseID-scenarioID}$ of root-causing premises (seeder).

To ensure the generality of our approach, we designed our framework to be compatible with several storage back-ends, including local file systems. This allows us to avoid the work involved with the task of uploading millions of files to a cluster of machines, which is time consuming (network bandwidth utilization) and consumes additional disk space (because of replication) [39]. We propose a custom hierarchical distributed framework for analytics, which is general enough to apply to similar analytical tasks. NAADSM generates variants that are already distributed across multiple machines, and we leverage an existing File Systems (FS) to work as Distributed File Systems (DFS), taking advantage of data locality. This system also provides check-pointing to relaunch an analysis task from where it is interrupted. Figure 4.3 shows an overview of the system orchestration used to meet our analytic requirements.

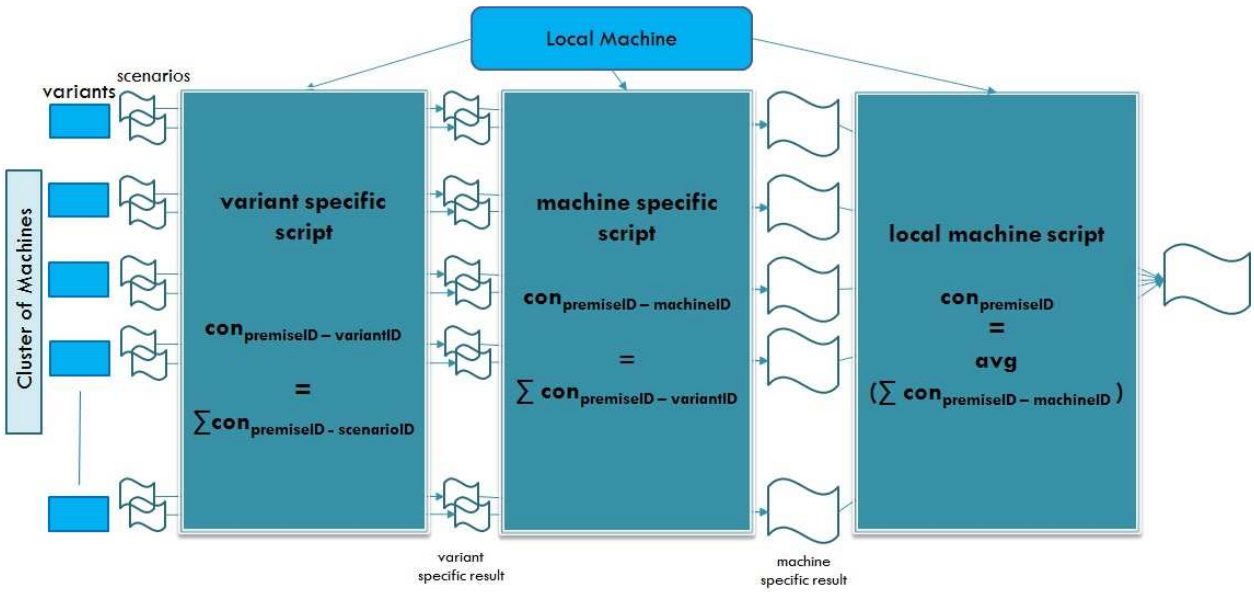



FIGURE 4.3. Overview of a distributed hierarchical aggregator used for empirical analytics.

Data is distributed over a cluster of machines, and from a local machine, we can launch variant-specific scripts and do calculations specific to each variant file containing multiple scenarios. In our analysis, we identify the scenario specific contribution of each premise and aggregate the results using variant-specific script. In this framework, we then can aggregate the results of all files located on associated machines using machine-specific scripts, launched from a local machine again. In the proposed approach, we combine premises' contributions to each variant. Then we can combine results from each machine and conclude the analysis on the local machine. With this methodology, we aggregate the results of all machines and take the average of the results to determine the premise-based contribution to the overall population. Further, we apply the 80-20 rule to select the top 20%. We observed that, for Colorado, the top 23.43% of infection contributors were responsible for 68.85% of the infections. This result provided a foundation for attribute-based modeling and classification.

## 4.2. Phase II: Model-Based Classification of Super-Spreaders

Super-spreaders behave differently from the rest of the population, and determining why a particular premise becomes a super-spreader can provide strong insight for disease spread analysis. Potential features that often influence super-spreaders include [18]:

- **Degree of Local Infections:** Number of units directly infected by a premise
- **Depth of Disease Transmission:** Length of the traversal path through the disease transmission network due to the associated premise's infection
- **Rate of Contribution:** Percentage of the total number of infected units by respective premise
- **Level at which Premise gets Infected:** Relative position of the premise in the infection chain

During the classification process, we backtrace through the disease spread network of each scenario to determine each of the above per-premise properties. There exist more than 3 million disease scenarios (∼100,000 variants * 32 iterations), and for each scenario each premise behaves differently. Therefore, we do not average the effect of each premise from each scenario because we want our model to be robust enough to the outliers as a reference point. Further, we have an accurate list of super-spreaders generated by empirical analysis. One of the most important goals of this thesis is to develop an analytic approach to a binary classification problem. We generated model that provides a binary justification for each premise according to its likelihood of being a super-spreader and our results supported highly accurate detection and interpretability of the model.

Consider the example scenario shown in Figure 4.4 regarding how we backtraced to each scenario and collected a feature set for each premise to create our machine learning dataset

buildup. We collected super-spreading premises using our empirical analytic approach (Section 4.1). For a given example scenario, the total number of infected premises are 10, and consider, for example, that B and F are super-spreaders in the population. The machine learning dataset collection may look as annotated in the Figure 4.4.

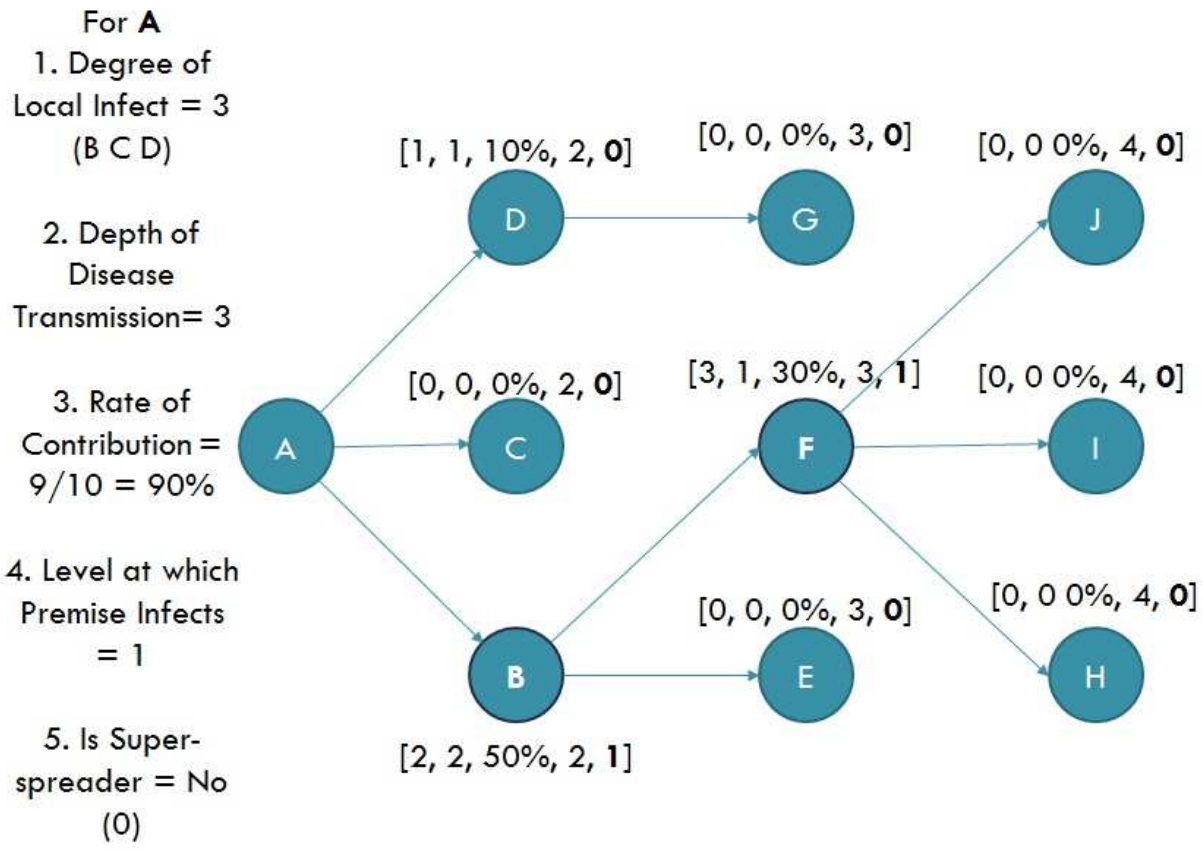



FIGURE 4.4. Disease spread scenario - disease spread network of example scenario.

Via backtracing, each premise-based property with binary justification was collected for all possible scenarios, and we turned the problem of pinpointing super-spreaders into a machine learning classification problem. As mentioned earlier, we had a voluminous dataset with millions of example sets with 4 feature sets per example. Considering the large training dataset, we applied a stochastic gradient descent methodology instead of batch (classic) gradient descent to maximize correct classification. Stochastic gradient descent (SGD) works

well with Big Data problems due to its scalability and fast convergence [13]. We applied several classifiers (models) with the help of SGD, including Support Vector Machine (SVM) and Logistic Regression with different regularization ($l_2$, *and* $l_1$ - *LASSO*) and tuned to achieve higher correct classification. An initial exploration of these models' hyper-parameters found that the classifications produced by Support Vector Machine applied with $l_2$ regularization exhibited the highest performance.

CHAPTER 5

# Evaluation and Validation

We leveraged the distributed hierarchical aggregation framework for our empirical analytics (Section 4.1) with scalable computing capabilities over a cluster of machines (∼10). The subject dataset for this thesis was generated by NAADSM (Colorado scenario). The dataset was distributed across a cluster and the framework took advantage of data locality for computation. The machines contained HP 4 core Xeon E3-1220 (3.1GHz) processors with 8GB RAM and 1TB memory disks. Our analysis result reported that, for Colorado, the top 23.43% of infection contributors were responsible for 68.85% of the infections. This result provided a foundation for our attribute-based classification. For our classification model, the training dataset was extracted using the same distributed framework . The dataset contained 3.2 million disease scenarios with an average of 6.65 infected premises per scenario. So ultimately, the dataset contained 21 million data points with 4 features (Section 4.2), and we applied various binary classifiers on it on local machine.

## 5.1. Classifying Super-Spreaders with Machine Learning

As described in Section 4.2, we turned the problem of super-spreader identification into a classification problem. Premises' classifications were stored in this dataset as binary values, with 1 indicating a super-spreader and 0 representing a regular herd. Our baseline classification via the 80-20 rule (empirical analytics —Section 4.1) was used as ground truth, and we followed SGD methodology with different models (i.e. SVM, Logistic Regression) and tuned parameters according to higher correct classifications.

Classifications were implemented with scikit-learn [40], and a randomized 70-30 split was used for the training and testing datasets, respectively. We applied SVM, Logistic Regression

with $l_2$ Regularization and $LASSO$ ($l_1$ Regularization). Using scikit-learn, we extracted an area under the curve score for different regularization terms, and graphical representation was as shown in Figure 5.1.

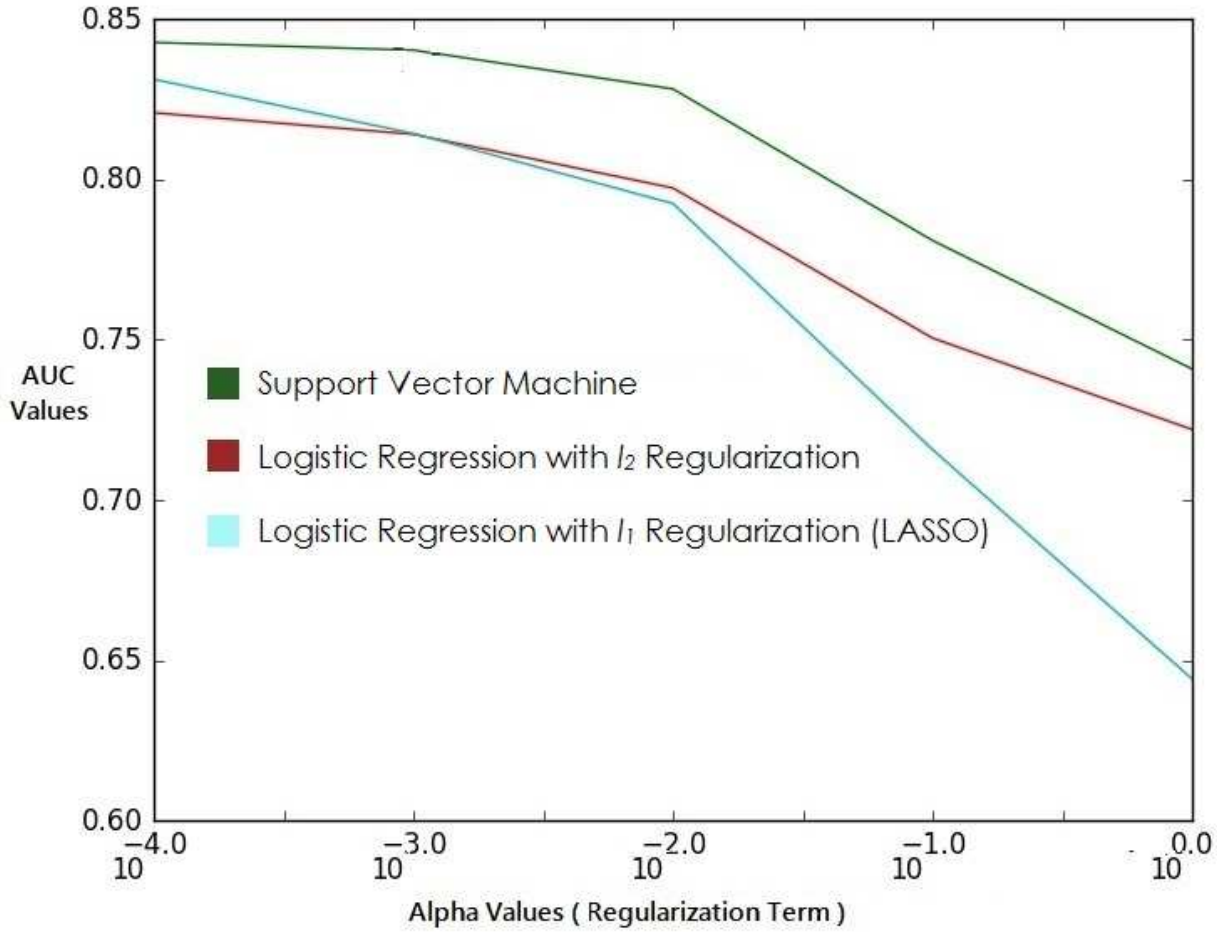



FIGURE 5.1. Area under the curve (AUC) for different models for test dataset of Colorado, USA.

As described in Figure 5.1, SVM worked best in all conditions and we chose it as our final model. SVM is also known as a large margin classifier and it tries to find the hyperplane that best represents the largest separation, or margin, between the two classes [41]. Accuracy was measured in terms of correct classification of normal premises and super-spreading premises overall. Accuracy was measured by the formula mentioned in equation 1.

$$(1) \qquad Accuracy = \text{Correct Classification of Premises} = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

Where,

TP = Number of classified super-spreading premises

FP = Number of normal premises misclassified as super-spreaders

FN = Number of super-spreaders misclassified as normal premises

TN = Number of correctly classified normal premises

TABLE 5.1. Machine Learning linear classifiers' evaluation for Colorado

| **Classifier (Applied Using SGD Methodology)** | **Accuracy** | **AUC Score** |
|---|---|---|
| Logistic Regression, $l_2$ Regularization, Regularization Parameter $\alpha = 0.0001$ | 89.10% | 0.8218 |
| Logistic Regression, $l_1$ Regularization - LASSO, $\alpha = 0.0001$ | 89.07% | 0.8203 |
| **Support Vector Machine, $l_2$ Regularization, $\alpha = 0.001$** | **89.97%** | **0.8458** |

As reported in Table 5.1, SVM is exhibiting better performance, however, it is worth noting that each of the machine learning algorithms (we applied linear classification methodology only) achieved reasonable accuracy based on our feature set, and the results agreed with the empirical analytic output.

One of the primary benefits of generating machine learning models is generalizability; if the model generalizes well, then it can predict super-spreaders in new or unseen datasets without the need to perform an analysis over the disease spreading network. To evaluate the generalizability of our SVM model trained on the Colorado dataset, we obtained a second

scenario set from Iowa, USA, which consisted of 8 TB of simulation output. Using the model, we were able to correctly classify premises with an accuracy of 93.5027%. Of this total, the correct classification rate of super-spreaders was 80.01%, and the correct classification rate of normal premises was 94.00%. The high level of accuracy in the correct classification of normal premises improves confidence that limited resources are not going to be wasted on normal premises. The Iowa dataset consists of more than 100 million data points, and the classification of all points took only 3.1591s.

### 5.2. Insights on the Relative Influence of Features

As reported in Section 5.1, SVM with SGD methodology achieved higher accuracy than other models in our case. After the algorithms are fully trained, coefficients associated with critical features capture the respective influence of each feature on classification [42]. We provide these coefficients as outputs during the modeling process. Coefficients from our SVM classifier are shown in Figure 5.2. Positive weights suggest a tendency for the classifier to classify a whole feature set as a super-spreader; whereas negative weights suggest otherwise. Based on these results, the degree of local infections exerts a strong influence with the premise in question being a super-spreader; this characteristic is also true of SARS outbreaks [30]. Conversely, unsurprisingly, the level at which a premise gets infected in the disease transmission hierarchy was negatively associated with being a super-spreader, and the contribution rate and depth of disease transmission were not weighted as highly for this particular model.

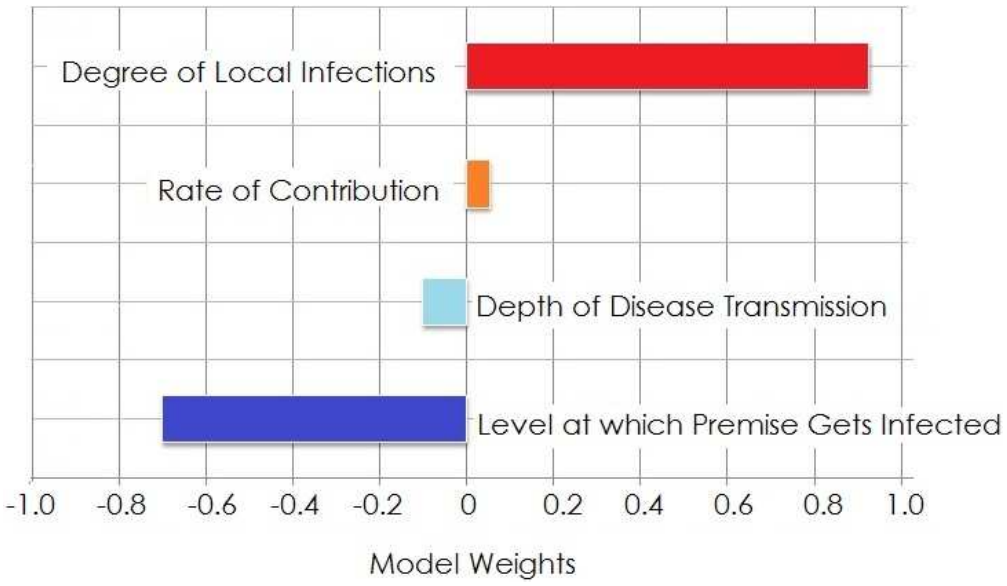



FIGURE 5.2. Feature coefficients from our Support Vector Machine classifier; larger values indicate more influential features.

The negligible influence of both **Rate of Contribution** and **Depth of Disease Transmission** makes our model more useful. The detection of both features requires us to see the whole disease spread chain, and it is not as useful to detect super-spreaders after disease outbreaks. In a given disease chain when particular premise is infected, based on its relative position in disease chain (i.e. level at which premise gets infected) and social connectivity (i.e. degree of local infections), our model can classify a premise as normal or as a super-spreader.

## 5.3. Comparison of Super-spreaders with Influential Premises found via PageRank Algorithm

To understand the composition of super-spreading premises, we applied a statistical technique on the data produced by our disease spread chain. Our analysis included ROC curves for the experiments. Practically, our hypothesis was that super-spreaders are the

most influential units in the disease spread network. We validated our hypothesis with the use of a list of highly influential premises found via the PageRank algorithm [15].

In this part of experiment, we analyzed the inclusion of super-spreaders in the composition of highly influential herds. For Colorado, we found 3747 probable super-spreaders (∼20% of 19000 total premises) using the approach described in Section 4.1. We then calculated the number of premises having the top n PageRank values among the 3747 super-spreaders, n ∈ {50, 100, 200, ..., 18800}. An overview of the experiment is diagrammed in Figure 5.3.

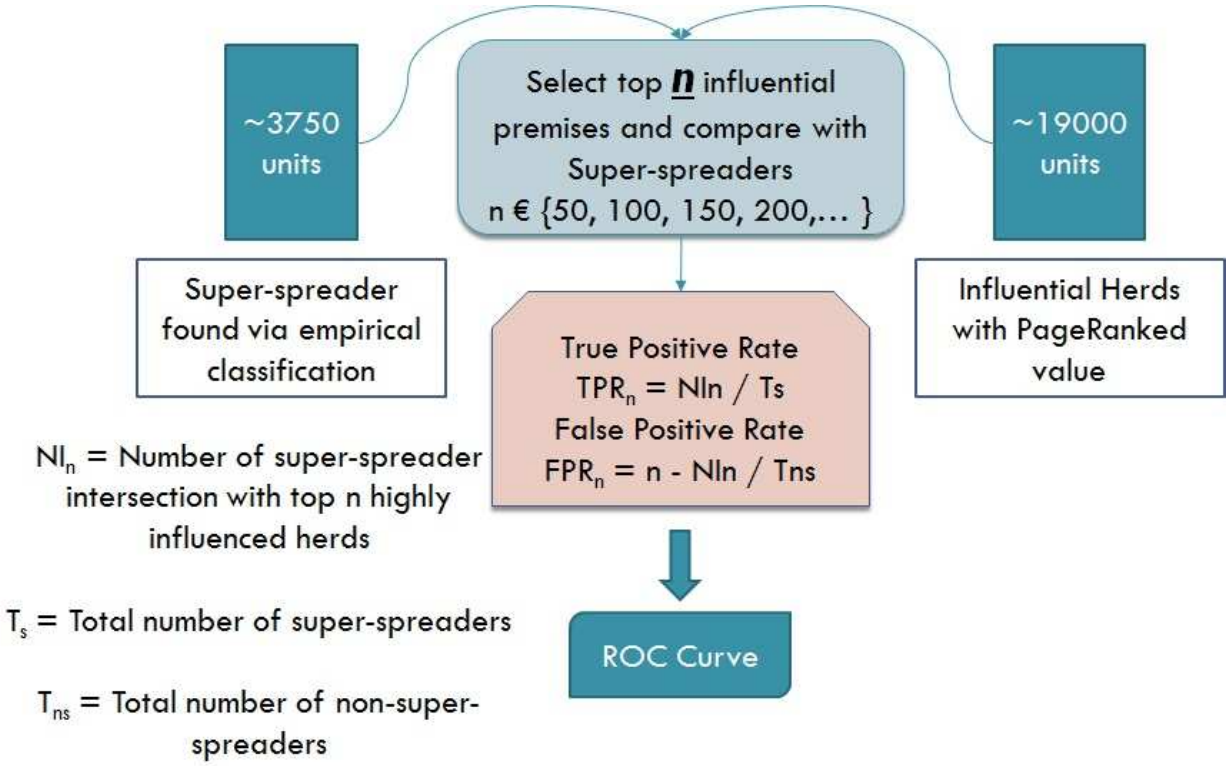



FIGURE 5.3. Abstract view of methodology for ROC curve generation.

The ROC curve for this experiment is shown in Figure 5.4.

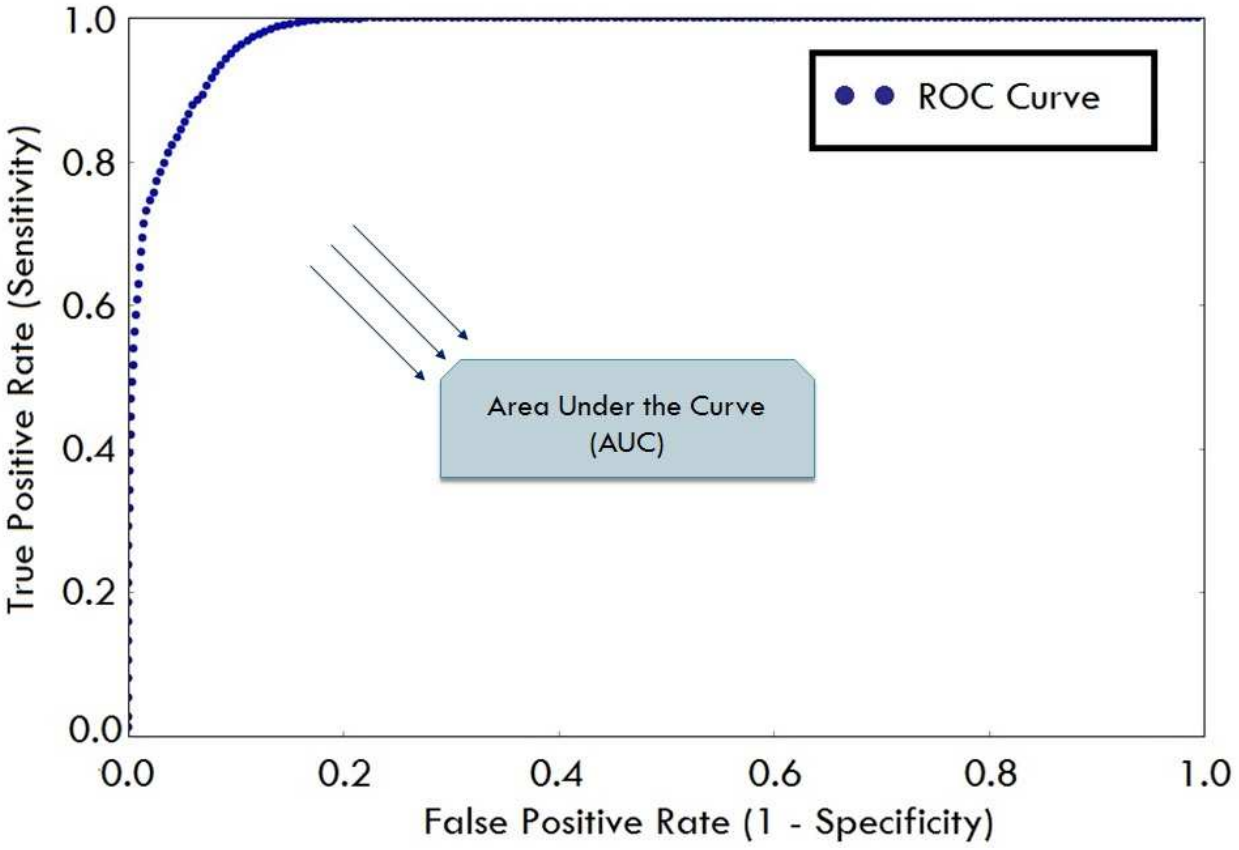



FIGURE 5.4. ROC curve for premises classified as super-spreaders compared with premises that exhibited high PageRank values (influential premises).

Based on the curve, the experiment resulted in high accuracy as the area under the curve ($\sim$0.9742) is reasonably high. These results indicate that super-spreaders account for a considerably large portion of the overall set of influential herds. The reason behind this result is that both groups infect a higher number of herds on average. According to Figure 5.2, the degree of local infection contributes most when classifying a herd as a super-spreader, and herds with high PageRank values tend to infect a higher number of herds overall. Moreover, we can observe that the likelihood ratio is decreasing as we move along the horizontal axis. The part of the curve with a high likelihood ratio refers to herds with high influence values, whereas the other part of the curve refers to the opposite.

## 5.4. Analyzing Geographic Location in Super-Spreading Events

SVM classifies premises based on feature sets in the positive region (super-spreaders) and the negative region (normal premises). The distance of a feature set from the classifier indicates confidence in the classification. Considering Figure 5.5, positive values that are larger (farther from the hyperplane) indicate super-spreaders with high confidence in classification, while larger negative values indicate normal herds with high confidence in classification. In both cases, values that are very close to the hyperplane represent weaker classifications.

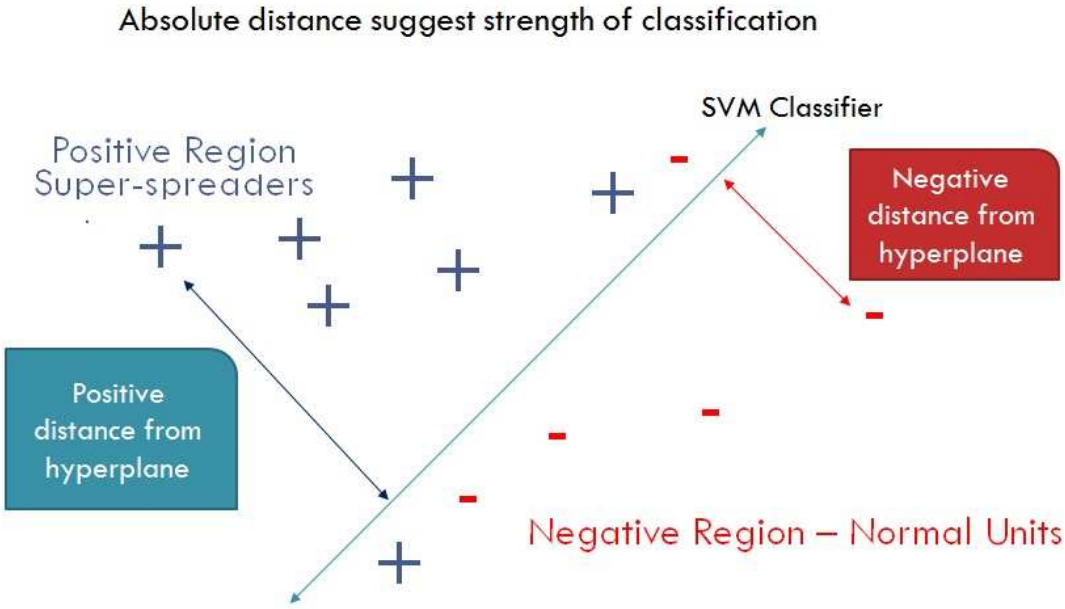
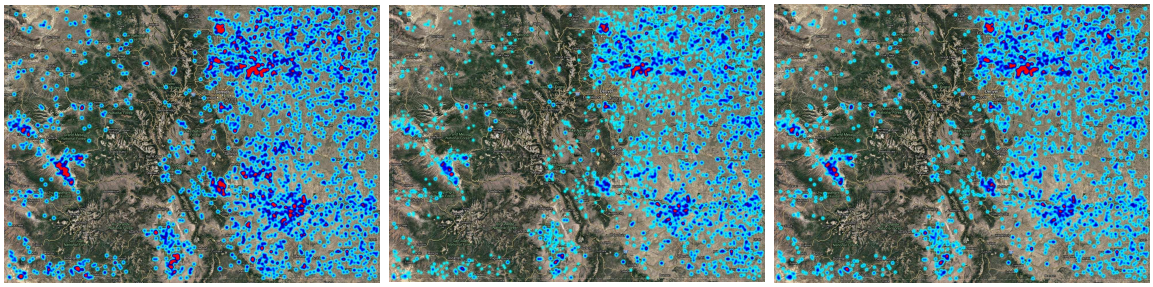



Figure 5.5. Support Vector Machine classification representation.

In Figure 5.6, we demonstrate the geographical distribution of premises in our Colorado dataset. Each graph contains a heat map depicting different approaches for classifying highly influential premises. Units with higher influence are highlighted by brighter shades of red, whereas less influential units are drawn in progressively darker shades of blue. These visualizations are based on the top 20% of the herds in the dataset to increase the level of contrast between premises. Three notable clusters can be seen in each of the subfigures, one in the mid-left, and another two near the top- and bottom right.



(A) The top 20% premises based on PageRank values.

(B) Premises based on their infection contribution towards population $cont_{premiseID}$.

(C) Super-spreaders based on their confidence in classification via our model.

FIGURE 5.6. Heat map of highly influential premises in Colorado, USA.

Figure 5.6a contains premise PageRank values, while the premise contribution to the overall infection ($cont_{premiseID}$) is shown in Figure 5.6b. The two heat maps are similar, indicating that the super-spreaders detected by premises contributions are a subcategory of the influential premises found via PageRank. On the other hand, Figure 5.6c represents the distance from the SVM classifier, which represents the confidence of the classification. Graphic comparison demonstrates the accuracy of our model; regions with high concentrations of super-spreaders and influential herds are classified with high confidence.

CHAPTER 6

# Conclusion and Future Work

In this study, we presented our methodology for identifying super-spreading premises and understanding their characteristics over voluminous data. Identification of such premises will help planners allocate limited resources more effectively and in a timely fashion. Our methodology accomplishes the identification of super-spreaders using an empirical analysis approach and classification from a voluminous dataset using machine learning technologies. In this thesis, we found that premises having higher degree of direct connection and exposure during the initial phase of a disease account for the greatest proportion of super-spreading premises. We validated our classification of premises with a link analysis algorithm (PageRank algorithm).

RQ1: Our statistical analysis demonstrates that super-spreaders are well-represented among highly influential premises. We have modeled the relationship between features of a premise extracted from the disease spread network and the likelihood of being a super-spreader using Supoort Vector Machines. Our model provides accuracy of 90% for simulated outbreaks in the state of Colorado; furthermore, this model transfers well and had an accuracy of over 93% when likely outbreaks in the state of Iowa were analyzed. This result supports the generalizability of our methodology.

RQ2: NAADSM has generated a voluminous dataset in order of TB and stored it in distributed machines in a Linux file system. So, instead of uploading a large dataset to HDFS and running a MapReduce job, we constructed a custom distributed hierarchical aggregation framework to run an empirical analysis. Although the framework

is not fault tolerant, we checkpoint its state to allow unfinished tasks to be relaunched. The distributed data scenario did not require iterative computation or inter node communication, and the framework used did an efficient job according to the empirical analysis that was conducted. Further, the classification model works efficiently even when dealing with more than 100 million data points, and it provides classification in less than 4 seconds because of the inherent properties of the stochastic gradient descent methodology.

While this thesis targets livestock disease outbreaks, the methodology that we describe is broadly applicable to systems where entities are organized into large networks and the spread of information (be it pathogens, ideas, or traffic movements) is based on relationships between entities.

As part of our future work we plan to add robustness and fault tolerance to our distributed framework. In addition, we would further like to explore feature space, such as geological location influence, to improve the accuracy of our super-spreader detection model. Another avenue for future research is to leverage input parameters that are used for simulation variants. This will include modeling the relationship between input features and super-spreaders.

# References


[1] E. Brooks-Pollock, M. de Jong, M. J. Keeling, D. Klinkenberg, and J. L. Wood, "Eight challenges in modelling infectious livestock diseases," *Epidemics*, vol. 10, pp. 1–5, 2015.

[2] P. D. Walsh, K. A. Abernethy, M. Bermejo, R. Beyers, P. De Wachter, M. E. Akou, B. Huijbregts, D. I. Mambounga, A. K. Toham, A. M. Kilbourn, *et al.*, "Catastrophic ape decline in western equatorial africa," *Nature*, vol. 422, no. 6932, pp. 611–614, 2003.

[3] M. Upton, "The role of livestock in economic development and poverty reduction," 2004.

[4] A. Estrada, R. Coates Estrada, D. Meritt Jr, D. Bojic Bultrini, A. Klein, I. Steffan Dewenter, T. Tscharntke, K. Taniguchi, X. Wang, P. Collomb, *et al.*, *World livestock 2011: livestock in food security.* No. FAO 363.8 W927 2011, FAO, Roma (Italia)., 2011.

[5] Wikipedia, "Zoonosis —wikipedia, the free encyclopedia," 2016. [Online; accessed 3-October-2016].

[6] J. K. Taubenberger and D. M. Morens, "1918 influenza: the mother of all pandemics," *Rev Biomed*, vol. 17, pp. 69–79, 2006.

[7] M. J. Keeling and P. Rohani, *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[8] N. Harvey, A. Reeves, M. A. Schoenbaum, F. J. Zagmutt-Vergara, C. Dubé, A. E. Hill, B. A. Corso, W. B. McNab, C. I. Cartwright, and M. D. Salman, "The north american animal disease spread model: A simulation model to assist decision making in evaluating animal disease incursions," *Preventive veterinary medicine*, vol. 82, no. 3, pp. 176–197, 2007.

[9] D. L. Pendell, J. Leatherman, T. C. Schroeder, and G. S. Alward, "The economic impacts of a foot-and-mouth disease outbreak: a regional analysis," *Journal of Agricultural and Applied Economics*, vol. 39, no. s1, pp. 19–33, 2007.

[10] C. Green, T. Whiting, G. Duizer, D. Douma, H. Kloeze, W. Lees, and A. Reeves, "Simulation modeling of alternative control strategies for an hpai outbreak using naadsm," in *Canadian Association of Veterinary Epidemiology Preventive Medicine (CAVEPM) Meeting*, 2010.

[11] K. Portacci, A. Reeves, B. Corso, and M. Salman, "Evaluation of vaccination strategies for an outbreak of pseudorabies virus in us commercial swine using the naadsm," *ISVEE*, vol. 12, p. 78, 2009.

[12] Wikipedia, "Pareto principle —wikipedia, the free encyclopedia," 2016. [Online; accessed 7-October-2016].

[13] T. S. Ferguson, "An inconsistent maximum likelihood estimate," *Journal of the American Statistical Association*, vol. 77, no. 380, pp. 831–834, 1982.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web.," 1999.

[15] N. Shah, H. Shah, M. Malensek, S. Pallickara, and S. Pallickara, "Network analysis for identifying and characterizing disease outbreak influence from voluminous epidemiology data," *(To Appear) IEEE International Conference on Big Data, Washington D.C., USA*, 2016.

[16] Z. Sui, N. Harvey, and S. Pallickara, "On the distributed orchestration of stochastic discrete event simulations," *Concurrency and Computation: Practice and Experience*, vol. 26, no. 11, pp. 1889–1907, 2014.

[17] Z. Sui, M. Malensek, N. Harvey, and S. Pallickara, "Autonomous orchestration of distributed discrete event simulations in the presence of resource uncertainty," *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, vol. 10, no. 3, p. 18, 2015.

[18] A. P. Galvani and R. M. May, "Epidemiology: dimensions of superspreading," *Nature*, vol. 438, no. 7066, pp. 293–295, 2005.

[19] Z. Shen, F. Ning, W. Zhou, X. He, C. Lin, D. P. Chin, Z. Zhu, and A. Schuchat, "Superspreading sars events, beijing, 2003," *Emerging infectious diseases*, vol. 10, no. 2, pp. 256–260, 2004.

[20] M. Malensek, W. Budgaga, S. Pallickara, N. Harvey, F. J. Breidt, and S. Pallickara, "Using distributed analytics to enable real-time exploration of discrete event simulations," in *Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference on*, pp. 49–58, IEEE, 2014.

[21] W. Budgaga, M. Malensek, S. Pallickara, N. Harvey, F. J. Breidt, and S. Pallickara, "Predictive analytics using statistical, learning, and ensemble methods to support real-time exploration of discrete event simulations," *Future Generation Computer Systems*, vol. 56, pp. 360–374, 2016.

[22] S. Funk, M. Salathé, and V. A. Jansen, "Modelling the influence of human behaviour on the spread of infectious diseases: a review," *Journal of the Royal Society Interface*, vol. 7, no. 50, pp. 1247–1256, 2010.

[23] S.-J. Paine, P. H. Gander, and N. Travier, "The epidemiology of morningness/eveningness: influence of age, gender, ethnicity, and socioeconomic factors in adults (30-49 years)," *Journal of biological rhythms*, vol. 21, no. 1, pp. 68–76, 2006.

[24] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, "Superspreading and the effect of individual variation on disease emergence," *Nature*, vol. 438, no. 7066, pp. 355–359, 2005.

[25] A. James, J. W. Pitchford, and M. J. Plank, "An event-based model of superspreading in epidemics," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 274, no. 1610, pp. 741–747, 2007.

[26] C. C. Aggarwal, A. Khan, and X. Yan, "On flow authority discovery in social networks.," in *SDM*, pp. 522–533, SIAM, 2011.

[27] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, ACM, 2003.

[28] B. Hajian and T. White, "Modelling influence in a social network: Metrics and evaluation," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 497–500, IEEE, 2011.

[29] A. A. Rad and M. Benyoucef, "Towards detecting influential users in social networks," in *International Conference on E-Technologies*, pp. 227–240, Springer, 2011.

[30] R. Fujie and T. Odagaki, "Effects of superspreaders in spread of epidemic," *Physica A: Statistical Mechanics and its Applications*, vol. 374, no. 2, pp. 843–852, 2007.

[31] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature physics*, vol. 6, no. 11, pp. 888–893, 2010.

[32] P. E. B. J. Feng, "Measuring user influence on twitter using modified k-shell decomposition," 2011.

[33] B. Xiang, Q. Liu, E. Chen, H. Xiong, Y. Zheng, and Y. Yang, "Pagerank with priors: An influence propagation perspective.," in *IJCAI*, Citeseer, 2013.

[34] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, pp. 261–270, ACM, 2010.

[35] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy.," *ICWSM*, vol. 10, no. 10-17, p. 30, 2010.

[36] A. Khrabrov and G. Cybenko, "Discovering influence in communication networks using dynamic graph analysis," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pp. 288–294, IEEE, 2010.

[37] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences of the United States of America*, pp. 16569–16572, 2005.

[38] M. Woolhouse, D. Shaw, L. Matthews, W.-C. Liu, D. Mellor, and M. Thomas, "Epidemiological implications of the contact network structure for cattle farms and the 20–80 rule," *Biology Letters*, vol. 1, no. 3, pp. 350–352, 2005.

[39] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*, pp. 1–10, IEEE, 2010.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[41] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[42] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.