Network Analysis for Identifying and Characterizing Disease Outbreak Influence from Voluminous Epidemiology Data

Naman Shah, Harshil Shah, Matthew Malensek, Sangmi Lee Pallickara and Shrideep Pallickara
Department of Computer Science
Colorado State University
{*namanrs, hkshah, malensek, sangmi, shrideep*}*@cs.colostate.edu*

***Abstract*—Planning for large-scale epidemiological outbreaks in livestock populations 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 that lead to extreme outbreaks. However, discovering and leveraging such information is also computationally expensive. In this study, we propose a distributed approach for analyzing voluminous epidemiology data to locate and classify the most influential entities in a disease outbreak. Using our *disease transmission network* (DTN), planners or analysts can isolate entities that have a disproportionate effect on epidemiological outcomes, enabling effective allocation of limited resources such as vaccinations and field personnel. We use a representative dataset to verify our approach, including identification of influential entities and creation of machine learning models for accurate classifications that generalize to other datasets.**

***Index Terms*—Epidemiological network analysis; Distributed analytics; Disease spread classification; Super-Spreading Events**

I. Introduction

According to the Food and Agricultural Organization (FAO), there are currently more than 1.5 billion cattle, 1.1 billion sheep, and 0.97 billion pigs and goats in the global livestock industry, which employs at least 1.3 billion people [1]. Effective planning and response to infectious threats in livestock are critical for the ecological system, the global economy, and human health in the case of zoonotic diseases (such swine flu) that exhibit cross-species transmission. There have been significant efforts in the epidemiological modeling community to understand and predict the distribution of disease within a herd as well as transmission between herds [2]. Epidemiological models, often expressed as stochastic discrete event simulations, involve hundreds to thousands of input parameters and tend to be compute-intensive.

In this study, we consider 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 [3]. NAADSM can be used to model foot-and-mouth disease (FMD), highly pathogenic avian influenza, swine flu, and pseudorabies [4], [5], [6]. In NAADSM, disease biology parameters include transmission via airborne or direct contact, control measures (such as vaccinations), effectiveness of vaccines, quarantines, shipments, and veterinarian visits. Since the simulation is stochastic, each set of input parameters is executed several times to gain statistical confidence in the results. These *iterations* contribute to the overall representation of the output variables' probability distributions. Key outputs used during planning include the disease duration, number of infected animals, and depletion of vaccine stockpiles. While this study 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.

One of the primary concerns during disease outbreak planning is allocating limited resources. Our goal in this effort is to identify premises that could contribute disproportionately to disease spread; i.e., once a particular premise is infected, the overall disease duration, total number of infections, and the probability of the disease becoming endemic are all high. Identifying such premises allows limited resources (vaccines, field personnel, and biosurveillance) to be allocated more effectively and in a targeted fashion. This involves analyzing voluminous data from simulation runs and tracking disease evolution over time. Pinpointing *highly influential herds* that contribute disproportionately to outbreaks is key when developing an effective response plan.

A. Scientific Challenges

Timely identification and characterization of influential herds introduces a set of unique challenges:

1) ***Dataset Size***: Epidemiological state is dispersed over a large number of files (3.2 million in our subject dataset). Each simulated time step produces an output file containing a variety of simulation data that must be processed to capture disease spread over time.
2) ***Timeliness***: Our algorithms and analysis workflows must execute in parallel across a cluster of computing resources to ensure timely results. Given the data volumes and disk I/O costs involved, repeated sweeps over the dataset would introduce significant delays in analysis.
3) ***Scalability***: The proposed approach must scale with increases in the number of premises and interconnectivity between entities. This ensures that the methodology is applicable in other scenarios.
4) ***Accuracy and Interpretability***: Our analysis must be reasonably accurate, and support interpretability by explaining why a herd is considered highly influential. This is critical for fine-tuning outbreak responses.

B. Research Questions

Research questions that we explore in this study include the following:

1) *What data structure(s) allow us to represent disease spread interactions for analysis?*
   Specifically, we must capture infection information from the simulation output and preserve the cumulative dynamics of disease spread. (§III-C)
2) *How can we measure the influence of each herd?*
   This involves discovering the epidemic characteristics of influential herds as well as the features that comprise these characteristics, which enables interpretability and herd classification. (§III-D)
3) *How can we enable the analysis at scale?*
   Given the data volumes involved, we must avoid repeated sweeps over on-disk data and execute analysis concurrently on multiple machines. Specifically, our methodology must scale with increases in the number of premises, contacts, and machines available for analysis. (§IV-D)

C. Overview of Approach

Our methodology for identification of influential premises in voluminous epidemiology data involves: (1) extracting relevant information needed for analysis from the dataset, (2) constructing a graph-based data structure, called the *disease transmission network* (DTN) to encode this information, (3) using the DTN for network analysis via the PageRank algorithm, and (4) identification and characterization of *super-spreaders* and *seeders*. Preprocessing and analysis tasks are expressed as distributed computations implemented using Apache Spark [7], with the dataset stored in HDFS [8]. These tasks execute concurrently on multiple machines with data locality, and avoid making repeated disk accesses by performing analysis in main memory.

Our epidemiology dataset encompasses multiple representative scenarios and iterations, which we process to extract and record millions of infection incidents. This includes tracking the number, source, destination, and duration of infections. This information is encoded in the disease transmission network. The DTN is a weighted, directed graph that summarizes the number of infections between premises; nodes within DTN are premises and edges represent infection transmissions. The direction of traversals within the DTN varies depending on the algorithm underpinning the analysis.

Once generated, we analyze the DTN in multiple steps to identify and characterize highly influential herds. One avenue we leverage for analysis is the PageRank algorithm, which was originally used in the Google search engine to estimate the importance of web pages [9]. In our study, we use PageRank to estimate the probability that a premise contributes to a random infection chain. We calculate PageRank values for each premise in the DTN; if a premise has a higher PageRank value, we consider the herd to be more influential in the disease outbreak.

Once we identify influential herds based on PageRank values, we perform further analysis to understand other epidemic characteristics such as classifying *super-spreaders* and *seeders*. In epidemiology, a super-spreader is a host that infects disproportionally more secondary contacts than other hosts. We use the Pareto Principle [10] to determine super-spreaders, and model the relationship between features extracted from the output dataset to classify the super-spreaders using support vector machines. On the other hand, seeders are hosts that are among the first to be infected. Besides global analysis using the DTN, we also allow identification of the most influential premise(s) on a local scale based on cross-premise reachability.

D. Paper Contributions

This paper presents our approach for identifying and characterizing highly influential herds by analyzing voluminous epidemiology data. Our specific contributions include:

1) We have designed a graph-based data structure, the *disease transmission network*, that preserves cumulative dynamics of disease spread across space and time. The data structure supports traversals that are needed for analysis and characterization.
2) Novel identification of influential herds by harnessing and adapting the PageRank algorithm in the context of epidemiology.
3) Support for interpretability of the analysis by identifying key features that characterize influential herds.
4) Classification of super-spreaders using support vector machines (SVMs). The resulting model can be used to inform why a particular premise should be given priority during outbreak responses.
5) Our approach avoids repeated I/O passes over the datasets and compactly encodes results in the memory-resident disease transmission network, which is amenable to subsequent analysis by multiple learning algorithms and statistical methods.

E. Paper Organization

The rest of the paper is organized as follows. Section II outlines the simulation and dataset used in this study, followed by our methodology in Section III. Subsection III-A describes the creation of the disease transmission network (DTN), followed by preliminary analysis in Subsection III-B. The remainder of our methodology is described in Subsection III-C, which describes how we identify influential entities in the DTN, and Subsection III-D, which details how we classify such entities. Section IV provides a thorough evaluation of our methodology, followed by related work in Section V. Finally, conclusions and future research directions are described in Section VI.

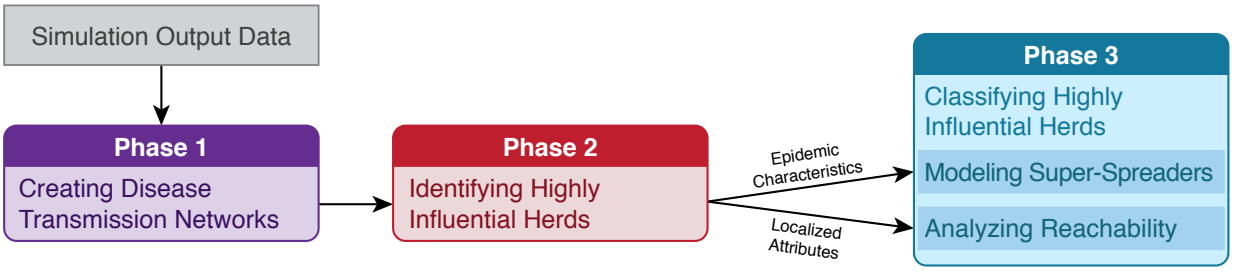


Fig. 1: High-level overview of our analysis workflow.

II. Background

A. NAADSM

The North American Animal Disease Spread Model (NAADSM) is a stochastic simulation of highly contagious disease outbreaks in animals to aid strategy development and decision making [3]. In this model, groups of livestock, called *units*, are the basis of the simulation. Note that we also use the terms *premise* and *herd* to refer to a group of animals. Disease spread between units is influenced by production types, inter-group similarities (shipment rates, infection rates, etc.), relative locations, and distances between herds. 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 and vaccination. Disease spread among units can happen in any of three methods: direct contact, indirect contact, and airborne spread. Stochastic processes drive all operations in the model and are based on user-defined distributions and relational functions. NAADSM input parameters can be of six types: yes/no values, integers, floating point numbers, probabilities, probability density functions, and relational functions. Collectively, these parameters form a *scenario*. Because the simulation is stochastic, it is generally run for several *iterations* (32 per scenario, in this study) to gain confidence in the output distributions. To reduce the overall execution time of the simulation, NAADSM can be parallelized over a cluster of computing resources in a fault-tolerant fashion [11].

B. 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 [12], [13]. This process generated 100,000 *scenario variants* that were executed 32 times for a total of 3.2 million outputs (6.26 TB). In this particular scenario, a single initial herd is infected, with disease spread eventually encompassing tens of thousands of premises. The output of the simulation contains attributes representing the disease status of individual premises (and their respective herds) 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.

C. System Components

We leverage the Spark framework [7] to provide scalable and fault-tolerant computing capabilities over a cluster of machines. Spark is used for writing applications to process large amounts of data which can be stored in distributed file systems (HDFS, S3), local file systems, or data streams, and includes functionality such as map, reduce, filter, and join. Compared to traditional MapReduce implementations, Spark allows in-memory, iterative computations. This is particularly beneficial for algorithms such as PageRank, and allows our analysis operations to avoid disk I/O unless absolutely necessary. We use Spark to generate disease transmission networks (DTNs) from our epidemiological simulation output dataset, as well as performing analysis of highly-influential herds based on the DTN. To facilitate distribution of files across the cluster and ensure data locality during computations, we use the Hadoop Distributed File System (HDFS) [8] to store our dataset and output files.

III. Methodology

In this study, our goal is to identify and classify highly influential herds in the disease outbreak network. To achieve this goal, we have composed a workflow that comprises multiple analysis phases. As depicted in Figure 1, there are 3 major phases. In Phase 1, we perform data preprocessing to extract features and create the disease transmission network that is leveraged by subsequent analysis steps. Phase 2 generates global herd rankings and influence measures from the DTN. Phase 3 focuses on characterizing highly influential herds by studying their epidemic attributes and modeling the relationship between the characteristics. We perform validation and evaluation for each phase in Section IV.

A. Creating DTNs

NAADSM generates one output file per scenario. Results for each iteration are assembled based on simulation time steps. A data fragment from an iteration contains over 2000 input variables and 10-20 output variables, including the outbreak duration, number of infected premises, and vaccinations used.

Since scanning the raw data for each analysis step is not efficient at this scale, we removed initially infected herds from each of 3.2 million iterations. With the remainder of the dataset, we generated a weighted directed graph called the *disease transmission network* (DTN). The DTN is denoted as $G = (V, E)$, where V is the set of vertices, representing herds, and E is the set of edges, representing infection propagation. To create the DTN, we extract *infection propagation pairs* from the dataset, which are tuples that include the infected herd and source of infection. We use the Spark framework to compute infection percentages for every infection propagation pair, which are used as the weights for directed edges in the graph. For example, if A and B are two vertices connected by an edge with weight $1/5$, then A is source of infection in 1 out of 5 instances where B is infected. Apart from removing initially infected herds, we did not perform additional pruning on the DTN because our methodology is robust to noise from low-impact entities in the source dataset.

B. Preliminary Analysis: Geospatial Distance

After our initial creation of the disease transmission network, we performed correlation analysis on the geospatial distance between units and rate at which a unit infects others. This evaluation served to test the functionality of the DTN as well as to gain insight as to how disease spread interactions behave spatially. Using the DTN, we calculated the infection rate between herds using following formula:

$$InfectionRate(A, B) = \frac{CountOfInfections(A, B)}{SourceOfInfection(A)} \quad (1)$$

Where:
$SourceOfInfection(A)$: total infections from unit A
$CountOfInfections(A, B)$: total infections that unit A transmitted to unit B.

The infection rate as defined in Formula 1 is calculated for every pair of herds in the DTN, as well as the geospatial distance between herds. Using these points of comparison, we calculated the Pearson Correlation Coefficient (PCC) for this data, which was -0.048, signaling that there is almost no correlation between the infection rate and distance between herds. This experiment demonstrates that with our particular scenario a diseased unit is no more likely to infect a herd in close proximity than those at greater spatial distances.

C. Identifying Highly Influential Herds

Influential herds play a pivotal role in transmitting disease to their neighbors by making outbreaks last longer or become more severe. In these situations, the influence of a unit depends on the influence of its neighbors. In other words, a unit has high influence if it is infecting other highly influential units. This type of interaction can be efficiently modeled by the PageRank algorithm.

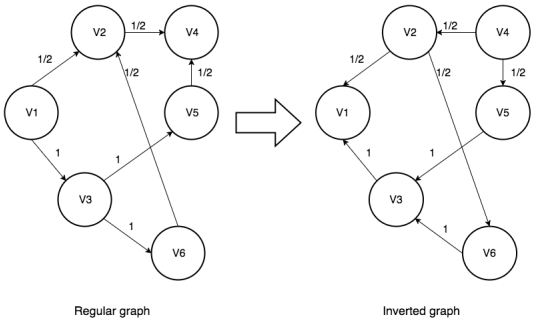


Fig. 2: Formation of an inverted graph of disease transmissions for use with the PageRank algorithm.

1) PageRank Algorithm: PageRank was proposed by Larry Page et al. [9] and used by the Google search engine to sort search results by their relevance or importance. The algorithm assigns a PageRank *value* to each web page, which describes the probability that a random surfer (randomly clicking on links) will arrive at the web page. The higher the PageRank value, more important the web page is. In general, highly linked pages are more important than pages with a low number of incoming links. Further, the PageRank value of a particular page determines how influential its outgoing links will be; if a page has very few input links but some are from highly linked web pages, then the page is ranked higher than a page that has more, but less important input links. This means that a website can achieve a high PageRank value either by having a large number of incoming links or by being linked to from an important page. This notion of importance is similar to being influential; considerable research has been conducted on using PageRank to determine influence [14], [15].

2) Using PageRank to Measure the Degree of Influence: Construction of the DTN produces a weighted, directed graph, where the weight of each edge is the rate at which one unit is infected by another. As a result, the sum of input links' weights must be equal to 1. When a disease is transmitted from vertex A to vertex B, we model the interaction as A *influencing* B. Similarly, vertex A influences all of its downstream neighbors. However, the PageRank algorithm computes the importance of entities based on *input* links, whereas in our case the influence of a vertex is decided by output links. Therefore, we invert the direction of edges in the graph without changing their weights to generate an *inverted* graph. This preserves the semantics of the network and allows usage of the PageRank algorithm without modification. A demonstration of an inverted graph is provided in Figure 2.

D. Classifying Highly Influential Herds

After discovering influential herds, we provide two types of classifications to understand their characteristics. First, we classify the herds based on their likelihood to be super-spreaders. Second, we perform *localized* classifications to detect herds that have a particularly strong influence on another herd but not necessarily the system as a whole.

In epidemiology, super-spreaders are a phenomenon that is widely observed in disease outbreaks. A super-spreader is an infected unit that spreads the disease disproportionally to other herds [16]. For a given outbreak, there may exist more than one super-spreader and the majority of individuals infect multiple secondary contacts. The most recent SARS outbreak involved super-spreading events (SSE) [17]. In this section, we investigate classifying super-spreaders from the group of highly influential herds. Classifying super-spreaders helps provide more efficient planning that controls contacts such as shipments or veterinarian visits.

1) Empirical Classification of Super-Spreaders: Super-spreaders tend to follow the Pareto principle [18], also known as the 80-20 rule, where approximately 20% of infected individuals are responsible for 80% of causality [10]. A herd is also considered to be a super-spreader if it is responsible for a significantly larger percentage of transmission [19]. To detect super-spreaders, we measure the per-herd *infection contribution* ($cont_{herdID}$) for each scenario by calculating the percentage of total infections caused by each herd. Infection contributions are collected from each scenario, averaged, and then sorted. We apply the 80-20 rule to select the top 20% of herds in descending order as probable super-spreaders, with all herds of equal ranking in the top 20% considered. Using this methodology, we observed that the top **23.43%** infection contributors were responsible for **68.85%** of the infections. This result provided a foundation for attribute-based modeling and classification.

2) Model-Based Classification of Super-Spreaders: Super-spreaders behave differently from the rest of the population, but determining *why* a particular herd becomes a super-spreader can provide high-level insight for disease spread analysis. Potential features that often influence super-spreaders include [16]:

- *Degree of local infections*: Number of units directly infected by a herd
- *Depth of disease transmission*: Length of the traversal path through the disease transmission network due to the associated herd's infection
- *Rate of contribution*: Percentage of the total number of infected units
- *Level of Infection*: Relative position of the premise in the infection chain hierarchy

We backtrace through the disease transmission network to determine each of these properties. After collecting training data for each herd across our subject dataset, we applied multiple machine learning classifiers: support vector machines (SVMs), random forests, and quadratic discriminant analysis (QDA). An initial exploration of these models' hyperparameters found that the classifications produced by SVMs exhibited the highest performance. To train the SVMs, we used *stochastic gradient descent* (SGD). SGD is a stochastic method for finding local minima or maxima by updating a set of parameters iteratively to minimize an objective function. The major advantage of SGD is its efficiency and amenability to parallel computation, which ensures scalability in our particular use case [20].

3) Reachability Analysis via Localized Attributes: Up to this point, discussion has revolved around determining influential herds across the entire disease transmission network. However, there are often localized relationships between herds that are significant but not highlighted by global analysis. Determining localized influence for a particular subset of herds is useful in situations where a planner wishes to isolate an infection or slow the spread of disease. These relationships are measured by the *localized* influence value, which is calculated based on Formula 2:

$$Infl\_val_{ij} = \frac{NPR_i * NOC_{ij}}{Avg\_dist_{ij}} \tag{2}$$

Where,
$Infl\_val_{ij}$= Influence value of herd *i* on herd *j*
NPR_i= Normalized PageRank value (1-10) of herd *i*, representing global influence in the DTN
NOC_{ij} = Normalized occurrence count (1-10) of herd *i* when herd *j* is infected
Avg_dist_{ij} is a measure of distance between herd *i* and herd *j*, which is calculated by the following formula:

$$Avg\_dist_{ij} = \frac{\sum_{k=1}^{n} dist_k(i,j)}{n} \tag{3}$$

Where,
n = Number of times herd *i* is infecting herd *j*
$dist_k(i,j)$ = Distance between herd *i* and herd *j* in hops for k^{th} occurrence

This results in herds having more influence on those in close proximity. For instance, a herd that is a single hop away is more influential than a herd that is two hops away in the DTN. Dividing NPR_i by Avg_dist_{ij} gives an approximate value of influence of herd *i* on herd *j*. By using NOC_{ij}, we increase the importance of herds that are infected often by a another herd.

IV. Evaluation

A. Experimental setup

The benchmarks and evaluations carried out in this study were performed on a cluster of 30 HP Z420 servers (8-core Xeon E5-2560V2, 32 GB RAM, 1 TB disk). Distributed computations were executed on Spark version 2.0 with the OpenJDK JVM, version 1.8.0_92. Each host was configured with Fedora 23 (Linux kernel 4.5.7). We used our epidemiological test dataset from Colorado, USA, which was distributed across the HDFS cluster (version 2.6.4), totaling 6.26 TB. Additional scenarios set in Iowa, USA, were used to verify the performance of our classifications, which consumed another 8.0 TB of disk space for a total dataset size of 14.26 TB.

B. Classifying Super-Spreaders with Machine Learning

Using the DTN to backtrace through herd interactions, we generated training data based on features that commonly indicate super-spreaders (as described in section III-D2). Herd

TABLE I: Accuracy for each machine learning classification algorithm evaluated. To demonstrate generality, we also used our SVM model on a different scenario set in Iowa, USA.

Classifier	Accuracy
Quadratic Discriminant Analysis	83.97%
Random Forest Classifier	88.9%
Support Vector Machine (SVM)	**90.02%**
SVM, Iowa Dataset	93.50%

classifications were stored in this dataset as a binary value, with 1 indicating a super-spreader and 0 representing a regular herd. Our baseline classification via the 80-20 rule was used as ground truth, and we applied several machine learning algorithms on the training data. Classifications were implemented with scikit-learn [20], and a randomized 90-10 split was used for the training and testing datasets, respectively. As depicted in Table I, the SVM model provided the highest accuracy. However, it is worth noting that each of the machine learning algorithms achieved reasonable accuracy based on our feature set.

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 needing to perform analysis over the disease transmission network. To evaluate the generality of our SVM model trained on the Colorado dataset, we obtained a second scenario set in Iowa, USA, which consisted of 8 TB of simulation output. Using the model, we were able to predict super-spreaders with an accuracy of 93.50% as shown in Table I. This is likely due to some similarities in parameters between the two scenarios, as both simulated an outbreak of foot-and-mouth disease.

After the algorithms are fully trained, coefficients associated with the features capture their respective impacts on classification. We provide these coefficients as outputs during the modeling process. Coefficients from our SVM classifier are shown in Figure 3; positive weights suggest a positive correlation with the output (classification as a super-spreader or not), and vice versa. Based on these results, the *degree of local infections* exhibits a strong correlation with the herd in question being a super-spreader, which is also true of SARS outbreaks [21]. Conversely, the level of infection in the DTN hierarchy was negatively correlated with being a super-spreader, and the contribution rate and depth of disease transmission were not weighted as highly for this particular model.

C. Statistical Evaluation of Super-Spreaders

To understand the composition of highly influential herds, we applied a variety of statistical techniques on the data produced by our disease transmission network. Our analysis includes a t-test, ROC curves for the experiments, as well as a breakdown of seeders, super-spreaders, and combined influential herds.

1) Highly Influential Herds vs Super-Spreaders: We performed a two-sample t-test to determine whether the tendency

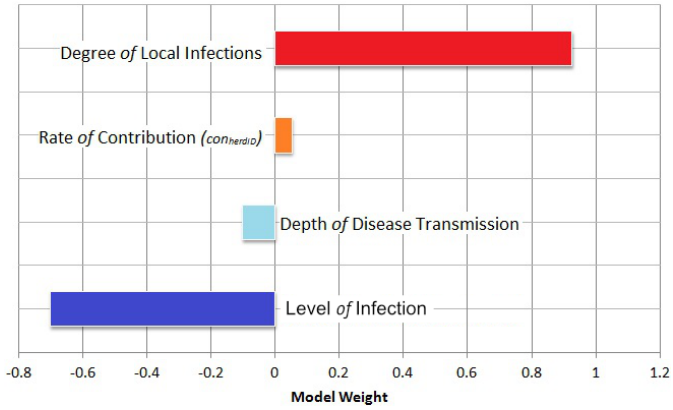


Fig. 3: Feature coefficients from our SVM classifier; larger values indicate more influential features.

to include super-spreaders in high- and low-PageRank herds was statistically significant. In this evaluation, we assessed the top 20% of PageRanked herds (likely super-spreaders) with the next 20%. To conduct the t-test, we generated 40 data points by randomly selecting 1000 herds from each set and noting the count of super-spreaders. This experiment revealed a significant difference between herds with high PageRank values ($\bar{x}_1 = 839.93, s_1 = 11.26$) and herds with low PageRank values ($\bar{x}_2 = 192.5, s_2 = 9.9$); $t(76.72) = 1.84, p = 0.03452$ for $\mu_0 = 643$. These results suggest that the mean number of super-spreaders found in both groups is notably different. Specifically, herds with high PageRank values contain 64.3% more super-spreaders.

In the next part of this experiment, we analyzed the inclusion of super-spreaders in the composition of highly influential herds. We found 3747 probable super-spreaders using the approach described in section III-D1. We then calculated the number of herds having the top n PageRank values among the 3747 super-spreaders, $n \in \{50, 100, 200, \cdots, 18800\}$. The ROC curve for this experiment is shown in Figure 4. Based on the curve, the experiment resulted in high accuracy, meaning 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 3, 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 as mentioned in III-C1. Moreover, we can observe that the likelihood ratio is decreasing as we move along horizontal axis. The part of curve with a high likelihood ratio refers to herds with high influence values, whereas the other part of the curve refers its counterpart.

2) Highly Influential Herds vs Seeders: This experiment analyzes the involvement of seeder herds (herds that are infected by the set of initially infected herds) in the evolution of super-spreaders. As described in Section III-A, we remove initially infected herds from the infection propagation pairs and collect the rest of the data for analysis. Over the 3.2 million iterations, we found 6504 distinct seeders. We per-

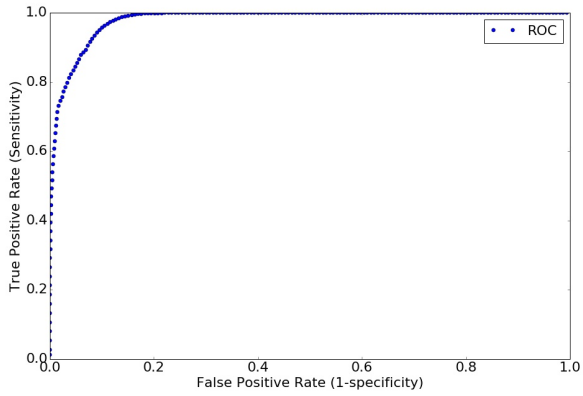


Fig. 4: ROC curve for herds classified as super-spreaders compared with herds that exhibited high PageRank values.

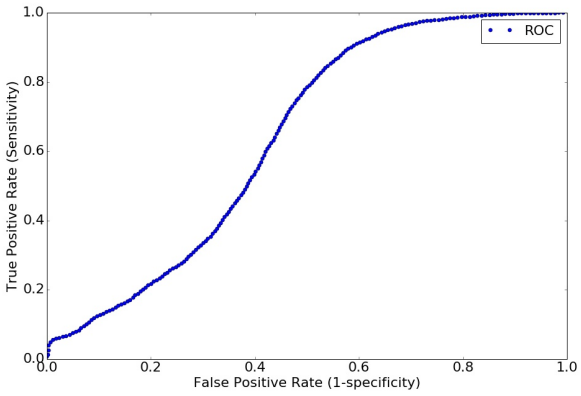


Fig. 5: ROC curve for herds classified as seeders compared with herds that exhibited high PageRank values.

formed same experiment as described in the previous section (IV-C1), except this time the number of herds having the top n PageRank value are among 6504 seeders instead of super-spreaders, $n \in \{50, 100, 200, \cdots, 18800\}$. The ROC curve for this experiment is shown in Figure 5; we can observe a small peak initially, followed by monotonic increases afterwards. The area under the curve is much less compared with the previous experiment performed on super-spreaders. This result suggests that seeders do not contribute to the composition of highly influential herds as much as the super-spreaders. There are likely two reasons for this: first, among the 6504 seeder herds, most are classified as seeders very few times in the overall dataset of 3.2 million simulated outbreaks, resulting in a lower number of overall infections. Second, seeders often infect herds with a low PageRank value, resulting in a little contribution towards their own influence.

The true Positive Rate (TPR) and False Positive Rate (FPR) used to create the ROC curves in the previous experiments are calculated using following formula:

$$TPR_n = \frac{NI_n}{T_p}$$
$$FPR_n = \frac{n - NI_n}{T_n} \quad (4)$$

Where:
NI_n = Intersection of super-spreaders or seeders with the top n highly influential herds
T_p = Total number of super-spreaders or seeders
T_n = Total number of non-super-spreaders or non-seeders

3) Highly Influential Herds vs the Union of Seeders and Super-Spreaders: To study the involvement of super-spreaders and seeders combined as a single group, we computed the union of the two sets to compare with highly influential herds derived from PageRank values. Figure 6 plots the size of each of these sets based on the top n PageRank values. This demonstrates that about 3000 of the top herds are either super-spreaders or seeders (with the majority being super-spreaders), as the initial portion of the curve overlaps with the identity line. After all the super-spreaders are accounted for (n=7100), the union set follows the shape of the seeder plot. This demonstrates that herds with the highest influence are largely super-spreaders.

D. Scalability Evaluation

We measured the time taken by the Spark framework to compute PageRank values of premises in the disease transmission network for various combinations of data and cluster sizes. From the 100,000 simulation outputs in our Colorado dataset, we extracted disease transmission information in the form of infection propagation pairs and executed our PageRank implementation. We considered cluster sizes with a varying number of nodes, each of which was accountable

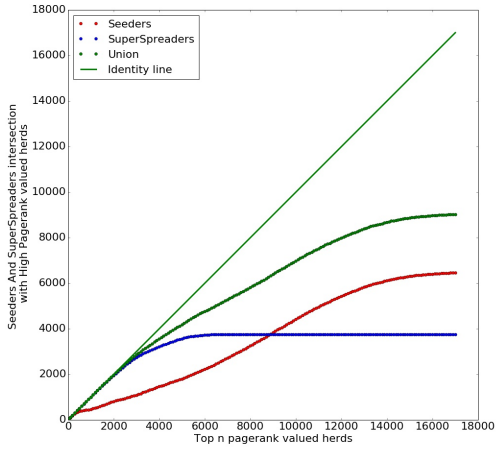


Fig. 6: Seeders, super-spreaders, and their union based on the top n PageRanked herds.

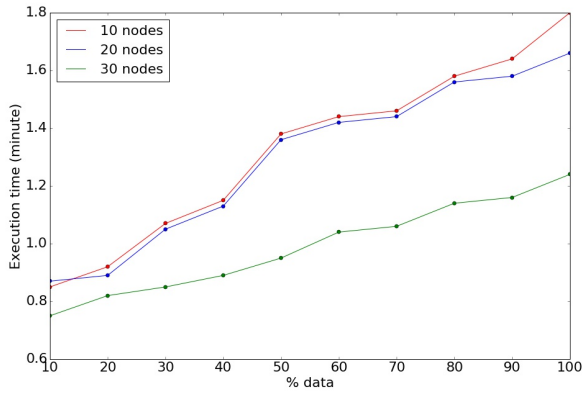


Fig. 7: Scalability of our approach executing under the Apache Spark framework. By increasing the cluster size to 30 nodes, we reduce the execution time by about 25%.

for four Spark workers. Figure 7 demonstrates the results of this benchmark; the vertical axis contains the time taken to perform the computation, with dataset sizes presented on the horizontal axis. Clusters of 10 and 20 machines exhibited similar execution times due to resource constraints that increased synchronization delays between stages, but the cluster of 30 machines improved computation times by about 25% for the full-sized dataset.

E. Analyzing Geographic Location in Super-Spreading Events

In Figure 8, we demonstrate the geographical distribution of herds in our Colorado dataset. Each graph contains a heat map depicting different approaches for classifying highly influential premises. Herds with higher influence are highlighted by brighter shades of red, whereas less influential herds are drawn in progressively darker shades of green. Note that 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.

Figure 8a contains herd PageRank values, while the premise contribution to the overall infection ($cont_{herdID}$) is shown in Figure 8b. Note that both heat maps are similar, indicating that the super-spreaders detected by herd contributions are a subcategory of the influential premises found via PageRank. On the other hand, Figure 8c depicts the distance from the hyperplane in our SVM classifier, which represents the confidence of the classification. Positive values that are larger (farther from the hyperplane) indicate super-spreaders with high confidence (shown in brighter red), while larger negative values indicate normal herds with high confidence (shown in darker green). In both cases, values that are very close to the hyperplane represent weak classifications.

As an alternative representation of this data, Table II contains the top 10 influential premises (by herd ID numbers) based on PageRank values, the contribution to the overall infection ($cont_{herdID}$), and distance from the hyperplane from our SVM model. Note that several of the premises appear in all three result sets.

TABLE II: Top premise IDs discovered by the approaches shown in Figure 8. Herds selected by multiple approaches are displayed in bold.

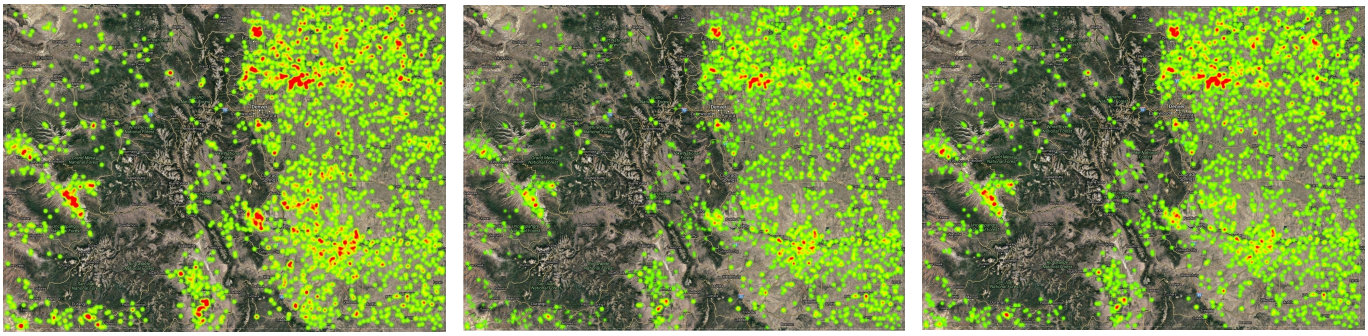
Top Premises Based On:		
PageRank Values	Contribution to Population	Distances from SVM Hyperplane
1220	**1683**	**11923**
1845	**1772**	**1845**
1683	**1620**	1052
1834	**1776**	1573
1914	17314	1074
1772	9825	16264
11923	1172	**1620**
1776	11241	11515
1913	1619	43
1837	**11923**	1894

V. Related Work

Influential herds transmit disease to their neighbors, ultimately making outbreaks last longer or become more severe. As a result, the influence of a herd depends largely on the influence of its neighbors. 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]. However, these approaches generally examine standalone characteristics and not the underlying network or relationships that result from disease spread.

Social Network Analysis (SNA) focuses on human interactions in social networks, but can be applied to analyze animal epidemics as well. Considerable research has been conducted on influence in social networks [14], [15], [24], [25], [26]. The Independent Cascade (IC) model and Linear Threshold (LT) model are commonly used to describe the influence of nodes in directed graphs. The LT model declares a node as either active or inactive based on a threshold and the sum of weights of neighboring edges. On the other hand, in the IC model, each active node is given an opportunity to activate its inactive neighbors, with the process repeating until a steady state is reached [27], [28]. In this case, active nodes are considered to be highly influential. However, since both of these methods rely on binary states (active or inactive), relative measures between nodes are not supported.

Cha et al. studies the influence of 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 [26]. An approach outlined by Khrabrov and Cybenko [29] uses daily mentions of users on Twitter as a basis for calculating different rank metrics such as PageRank, drank, and StarRank to determine influence.



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

(b) Premise contributions towards the overall infection ($cont_{herdID}$).

(c) Super-spreader classifications using our SVM machine learning model.

Fig. 8: Heat map of highly influential premises in our Colorado dataset.

Aggarwal et al. [15] proposes two algorithms, SteadyStateSpread and RankedReplace, to determine *information flow representatives*, a small group of authoritative figures to whom the release of information leads to maximum spread. SteadyStateSpread iteratively finds a candidate set of nodes with higher steady state *flow values* as candidate representatives. This method ignores the structural relationship of nodes, which inspired the RankedReplace algorithm. In RankedReplace, nodes are replaced iteratively and sorted in descending order by their steady state flow values to maximize total flow [15].

Substantial effort has been devoted to identifying hotspots that result in super-spreading events (SSEs). Lloyd-Smith et al. defines a protocol to identify super-spreaders, which is applicable in understanding SARS outbreaks [19]. The protocol suggests that the mean number of secondary infections from a particular host follows a Poisson distribution and outliers are often accountable for super-spreading events. However, underestimation of the epidemic potential can occur when field observations of mean secondary infections are low [30]. Fujie-Odagaki et al. focuses on intrinsically strong herd infectiousness and social connections [21]. Our particular dataset, however, does not reveal such information.

Epidemiological big data analysis systems include Google Flu Trends [31], which uses web search data to model flu-like symptoms in user queries and leverages the correlation between medical searches and physician visits to estimate influenza activity across the United States. The system provides results faster than traditional disease surveillance methods, and aids in the prediction and mitigation of seasonal influenza epidemics. Galileo [32], [33], [34], [35] uses a graph-based indexing scheme to enable analysis between entities in multidimensional data, with support for spatial queries based on proximity, polygons, or administrative boundaries [36]. SWAN [37] is a distributed knowledgebase for coordinating and researching Alzheimer Disease. By using semantic web concepts and variable privacy settings, researchers can collect information and collaborate while also avoiding duplicated effort. While SWAN handles data management, analytics activities must be carried out using other software packages.

VI. Conclusions and Future Work

In this study, we presented our methodology for identifying epidemiologically influential premises and understanding their characteristics over voluminous data. Identification of influential premises will help planners allocate limited resources more effectively. Our methodology includes multiple analysis components such as: (1) generating a disease network data structure, (2) estimating the influence of a particular premise using the PageRank algorithm, and (3) characterizing influential premises based on their epidemiological characteristics and premise-based relevance.

RQ1: To achieve effective analysis with reasonable latency, we extract entire chains of infections from the output dataset and construct a graph-based disease transmission network (DTN) that represents a holistic view of disease transmissions by maintaining the probability of infections between each herd pair. The DTN is a compact data structure that is less than 0.002% of the original dataset size. Since infections between herds are observed over 3.2 million iteration outputs, maintaining this pairwise probability with the DTN reduces the number of I/O accesses (encompassing both disk and network I/O) to the dataset significantly.

RQ2: We leverage the PageRank algorithm to estimate the influence of each herd in the DTN. The PageRank associated with a premise represents the probability that it contributes to a random infection chain. Our statistical analysis demonstrates that super-spreaders are well-represented among the highly influential premises. We have modeled the relationship between features of a premise extracted from the DTN and the likelihood of being a super-spreader using support vector machines (SVMs). Our model provides an accuracy of greater than 90% for FMD outbreaks in the state of Colorado; furthermore, this model transfers well and has an accuracy of over 93% when analyzing likely outbreaks in Iowa. This result demonstrates the generalizability of our methodology.

RQ3: Our analysis and experiments were performed using Apache Spark and were distributed across a cluster of computing resources. This approach was shown to be effective and scalable in our benchmark evaluation.

As part of our future work we plan to explore the feature

space to improve the accuracy of our super-spreader detection model. We will extend the DTN data structure to include other features such as types of premises. Another avenue for future research is to leverage input parameters that are used for simulation variants to model the relationship between input features and highly influential premises.

ACKNOWLEDGMENT

This work was supported by the US Department of Homeland Security [HSHQDC-13-C-B0018, D15PC00279]; and the US National Science Foundation's Advanced Cyberinfrastructure and Computer Systems Research Programs [ACI-1553685, CNS-1253908].

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

[3] N. Harvey, A. Reeves, M. A. Schoenbaum *et al.*, "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.

[4] D. Pendell, J. Leatherman, T. Schroeder, and G. Alward, "The economic impacts of a foot-and-mouth disease outbreak: a regional analysis," *Journal of Agricultural and Applied Economics*, vol. 39, no. 0, pp. 19–33, 2007.

[5] 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, May 29 - 30 2010, Guelph, Ontario, Canada*, 2010.

[6] 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," in *ISVEE 12: Proceedings of the 12th Symposium of the International Society for Veterinary Epidemiology and Economics, Durban, South Africa*, 2009, p. 78.

[7] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, ser. HotCloud'10, 2010, pp. 10–10.

[8] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop distributed file system," in *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, ser. MSST '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 1–10.

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

[10] Wikipedia, "Pareto principle — wikipedia, the free encyclopedia," 2016, [Online; accessed 25-July-2016]. [Online]. Available: \url{https://en.wikipedia.org/w/index.php?title=Pareto_principle&oldid=731439344}

[11] Z. Sui, M. Malensek, N. Harvey, and S. Pallickara, "Autonomous orchestration of distributed discrete event simulations in the presence of resource uncertainty," *ACM Trans. Auton. Adapt. Syst.*, vol. 10, no. 3, pp. 18:1–18:20, Sep. 2015.

[12] W. Budgaga, M. Malensek, S. L. 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 Gener. Comput. Syst.*, vol. 56, no. C, pp. 360–374, Mar. 2016.

[13] M. Malensek, W. Budgaga, S. L. Pallickara, N. Harvey, F. J. Breidt, and S. Pallickara, "Using distributed analytics to enable real-time exploration of discrete event simulations," in *Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, ser. UCC '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 49–58.

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

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

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

[17] 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.

[18] 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.

[19] 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.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] 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.

[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] 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*. ACM, 2003, pp. 137–146.

[25] 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*. IEEE, 2011, pp. 497–500.

[26] 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.

[27] J. Goldenberg, B. Libai, and E. Muller, "Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata," *Academy of Marketing Science Review*, vol. 2001, p. 1, 2001.

[28] ——, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing letters*, vol. 12, no. 3, pp. 211–223, 2001.

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

[30] 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.

[31] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[32] M. Malensek, S. L. Pallickara, and S. Pallickara, "Autonomous cloud federation for high-throughput queries over voluminous datasets," *IEEE Cloud Computing*, vol. 3, no. 3, pp. 40–49, May 2016.

[33] ——, "Analytic queries over geospatial time-series data using distributed hash tables," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 6, pp. 1408–1422, June 2016.

[34] ——, "Fast, ad hoc query evaluations over multidimensional geospatial datasets," *IEEE Transactions on Cloud Computing*, p. (To Appear).

[35] C. Tolooee, M. Malensek, and S. L. Pallickara, "A scalable framework for continuous query evaluations over multidimensional, scientific datasets," *Concurrency and Computation: Practice and Experience*, vol. 28, no. 8, pp. 2546–2563, 2016, cpe.3651.

[36] M. Malensek, S. L. Pallickara, and S. Pallickara, "Evaluating geospatial geometry and proximity queries using distributed hash tables," *Computing in Science & Engineering*, vol. 16, no. 4, pp. 53–61, 2014.

[37] Y. Gao, J. Kinoshita, E. Wu, E. Miller, R. Lee, A. Seaborne, S. Cayzer, and T. Clark, "Swan: A distributed knowledge infrastructure for alzheimer disease research," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 4, no. 3, pp. 222–228, 2006.