KDD 2018: The 24th ACM SIGKDD International
Conference on Knowledge Discovery and Data Mining

epiDAMIK: Epidemiology meets Data Mining and Knowledge discovery

Workshop held in conjuction with ACM SIGKDD 2018

London, UK. August 20, 2018

Workshop Proceedings

Editors: B. Aditya Prakash, Anil Vullikanti, Shweta Bansal, Adam Sadelik

Proceedings of the ACM SIGKDD Workshop on Epidemiology meets Data Mining and Knowledge Discovery (epiDAMIK)

epiDAMIK'18, August 20, 2018, London, UK.

ACM SIGKDD Workshop on Epidemiology meets Data Mining and Knowledge Discovery (epiDAMIK)

Organizers:

B. Aditya Prakash (Virginia Tech)
Anil Vullikanti (Virginia Tech)
Shweta Bansal (George Washington University)
Adam Sadelik (Google)

Program Committee:

Jose Cadena (Lawrence Livermore National Laboratory)
Ambuj Singh (University of California, Santa Barbara)
Ingmar Weber (Qatar Computing Research Institute)
Arvind Ramanathan (Oak Ridge National Laboratory)
Victor Preciado (University of Pennsylvania)
Jilles Vreeken (Max-Planck Institute for Informatics and Saarland University)
Nikhita Vedula (The Ohio State University)
Daniel Neil (Carnegie Mellon University/New York University)
Kathryn Rough (Google)
Bryan Lewis (Virginia Tech)

Webmaster:

Bijaya Adhikari (Virginia Tech)

Preface

With increasing globalization, urbanization, and ecological pressures, the threat of devastating global pandemics becomes more pronounced. The impact of Zika, MERS, and Ebola outbreaks over the past decade has strongly illustrated our enormous vulnerability to emerging infectious diseases. There is an urgent need to develop sound theoretical principles and transformative computational approaches that will allow us to address the escalating threat of a future pandemic. Data mining and Knowledge discovery have an important role to play in this regard. Different aspects of infectious disease modeling, analysis and control have traditionally been studied within the confines of individual disciplines, such as mathematical epidemiology and public health, and data mining and machine learning. Coupled with increasing data generation across multiple domains (like electronic medical records and social media), there is a clear need for analyzing them to inform public health policies and outcomes. Recent advances in disease surveillance and forecasting, and initiatives such as the CDC Flu Challenge, have brought these disciplines closer––public health practitioners seek to use novel datasets and techniques whereas researchers from data mining and machine learning develop novel tools for solving many fundamental problems in the public health policy planning process. We believe the next stage of advances will result from closer collaborations between these two communities, which is the main objective of epiDAMIK. The workshop is also an integral part of the 'Health Day @ KDD' this year, which is bringing together domain and machine learning experts to discuss challenges and trends in the healthcare industry as well as techniques and methodologies the machine learning community is using, and in process of developing to address these challenges.

This year's flu season has been severe (by some accounts, the worst since 2009), resulting in many lives lost and economic damages. This coincides with the 100th year anniversary of the worst pandemic in human history (the 1918 influenza pandemic). Hence the impact of infectious disease epidemics has been in sharp focus worldwide with an intense interest in real progress in this area from the public, government and academic stakeholders.

The main program of epiDAMIK'18 consists of nine papers that cover various aspects of data mining and public health. In addition there were two keynotes. Two papers were presented orally, and seven were presented during the interactive poster session. In addition, five papers were also jointly presented at the KDD poster session as part of the Health Day. These papers were selected after a thorough reviewing process. We sincerely thank the authors of the submissions and the attendees of the workshop. We also wish to thank the members of our program committee for their help in selecting a set of high-quality papers. Furthermore, we are very grateful to Rumi Chunara and Madhav Marathe for engaging keynote presentations.

B. Aditya Prakash
Anil Vullikanti
Shweta Bansal
Adam Sadelik

Blacksburg, August 2018

Table of Contents

Invited Talks

Data Mining and Machine Learning For Public Health 2.0

Rumi Chunara 7

Towards real-time epidemic science: Leveraging scalable computing, AI and data science

Madhav Marathe 8

Research Papers

Repeated Active Screening of Networks for Diseases

Biswarup Bhattacharya, Han Ching Ou, Arunesh Sinha, Sze-Chuan Suen, Bistra Dilkina and Milind Tambe 9

Validation of Network-Dependent Epidemic Processes: A Study of Dr. Snow's Seminal Cholera Dataset

Philip Pare, Ji Liu, Carolyn L. Beck, Tamer Basar and Angelia Nedich 17

Forecasting the Flu: Designing Social Network Sensors for Epidemics

Huijuan Shao, K.S.M. Tozammel Hossain, Hao Wu, Maleq Khan, Anil Vullikanti, B. Aditya Prakash, Madhav Marathe and Naren Ramakrishnan 21

A Topological Data Analysis Approach to Influenza-Like-Illness

Joao Pita Costa, Primož Škraba, Daniela Paolotti and Ricardo Mexia 29

Critical spatial clusters for vaccine preventable diseases

Jose Cadena, Achla Marathe and Anil Kumar Vullikanti 33

Identification of at risk groups for opioid addiction through web data analysis

Kaustav Basu, Sandipan Choudhuri, Arunabha Sen, Aniket Majumdar and Dipak Dey . . 41

Epidemiological Data and Model Requirements to Support Policy

Marc Baguelin, Elizabeth Buckingham-Jeffery, Ian Hall, Thomas House, Timothy Kinyanjui and Lorenzo Pellis 45

Dynamics underlying global spread of emerging epidemics: An analytical framework

Lin Wang and Joseph Wu 50

Cleanliness Campaign V/S Sanitation Related Diseases - Are they parallel in public perspective?

Aarzoo Dhiman, Durga Toshniwal, Soumya Somani, Preeti Malik 59

Invited Talk

Data Mining and Machine Learning For Public Health 2.0

Rumi Chunara
Assistant Professor
Department of Computer Science & Engineering
and College of Global Public Health
New York University
rumi.chunara@nyu.edu

Abstract:
High-resolution data sources can improve how we target intervention efforts and spend public health budgets. However, if the data is unstructured or generated observationally by select populations, there are computational challenges that must be addressed in order to assess practical epidemiological measures from it or include the data in epidemiological models. In this talk I will discuss areas in which we are addressing these challenges by using data mining and machine learning approaches to generate high-resolution features for epidemiological models and improving prediction efforts in both infectious and non-communicable diseases. Examples will include domain adaptation for predicting infection from community-sourced data, natural language processing to learn temporal representations of health behaviors, and unsupervised methods to learn spatial representations of health risks from noisy, irregular and sparse data.

Bio:
Rumi Chunara is an Assistant Professor at NYU, jointly appointed at the Tandon School of Engineering (in Computer Science) and the College of Global Public Health. Her research interests are in using person-generated data sources for population-level disease surveillance. In doing so, Dr. Chunara develops statistical and machine learning methodology for using these observational data sources in disease modeling efforts. Dr. Chunara joined NYU in 2015. Previously she was an Instructor at Harvard Medical School, the Children's Hospital Informatics Program and HealthMap. She completed her PhD at the Harvard-MIT Division of Health, Sciences and Technology, Master's degree at MIT in Electrical Engineering and Computer Science and received her Bachelor's degree in Electrical Engineering from Caltech with honors. Her research has been reported on widely including in NBC.com, CNN.com, and Scientific American. Chunara is a recipient of the MIT Presidential Fellowship and a Caltech Merit Scholarship, the NYC Media Lab - Bloomberg Data for Good Exchange Paper Award and was selected as an MIT Technology Review Top 35 Innovator under 35 in 2014.

Invited Talk

Data Mining and Machine Learning For Public Health 2.0

Madhav Marathe
Professor
Biocomplexity Institute
and Department of Computer Science
Virginia Tech.
mmarathe@bi.vt.edu

Abstract:
The H1N1 pandemic of 2009 and the 2014 Ebola outbreak in West Africa serve as a reminder of the social, economic and health burden of infectious diseases. The ongoing trends towards urbanization, global travel, climate change and a generally older and immuno-compromised population continue to make epidemic planning and control challenging. Recent advances in computing, AI and bigdata have created new opportunities for realizing the vision of real-time epidemic science. In this talk I will overview of the state of the art in computational epidemiology with an emphasis on computational thinking and on the development of scalable and pervasive computing techniques for planning, forecasting and response in the event of epidemics. I will draw on our work in supporting federal agencies during recent epidemic outbreaks as well as the development of epidemic forecasting methods in collaboration with federal and commercial partners. Computational challenges and directions for future research will be discussed.

Bio:
Madhav Marathe is the director of the Network Dynamics and Simulation Science Laboratory, Biocomplexity Institute and a professor in the department of computer science, Virginia Tech. His research interests are in network science, foundations of computing, high performance computing and their applications to problems in one-health, inter-dependent infrastructures and social sciences. Before coming to Virginia Tech, he was a Team Leader in the Basic and Applied Simulation Science Group that is a part of the Computer and Computational Sciences division at the Los Alamos National Laboratory (LANL). He is a Fellow of the IEEE, ACM, SIAM and AAAS.

Repeated Active Screening of Networks for Diseases

Biswarup Bhattacharya, Han Ching Ou
University of Southern California
Los Angeles, California, USA
bbhattac@usc.edu

Arunesh Sinha
University of Michigan
Ann-Arbor, Michigan, USA

Sze-Chuan Suen, Bistra Dilkina, Milind Tambe
University of Southern California
Los Angeles, California, USA

ABSTRACT

An important means of controlling recurrent infectious diseases is through active screening to detect and treat patients. Disease detection on a large network of individuals is a challenging problem, as the health states of individuals are uncertain and the scale of the problem renders traditional dynamic optimization models impractical. Moreover, efficient use of diagnostic and labor resources is a major concern, especially when the recurrent disease is prevalent in a resource-constrained region. In this paper, we propose a novel active screening model and an algorithm to facilitate active screening for recurrent diseases. Our contributions include: (1) A new approach for modeling SEIS type diseases using a novel belief-state representation, (2) a community and eigenvalue-based algorithm (TRACE) to perform multi-round active screening. We perform extensive experiments on real-world datasets which emulate human contact, and illustrate significant benefits due to TRACE.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent planning**; **Partially-observable Markov decision processes**; • **Applied computing** → **Life and medical sciences**;

KEYWORDS

Public health; SEIS Disease Model; Active Screening; Eigenvalue; Community; Belief states

1 INTRODUCTION

Curable infectious diseases are responsible for millions of deaths every year. Tuberculosis (TB), one such disease, affected over 10 million people worldwide in 2016, and caused over 400,000 deaths in India, the country with the highest TB mortality [28]. While low-cost treatment programs are available, many rely on patients to seek medical care (*passive screening*). However, individuals mistake their symptoms for another condition and not seek care. Public health agencies therefore engage in *active screening*, where individuals in the community are asked to undergo diagnostic tests and are offered treatment if tests return positive results [16].

It is costly to seek out at-risk individuals, and active screening efforts are often limited to high risk groups such as household TB contacts [9]. This method can successfully identify patients [3], and has been extensively evaluated [17]. However, this approach can be challenging to implement widely in resource-constrained regions such as India, as there are large transmission networks of potential patients and the number of health workers is limited. Prior studies show that even when focusing on high-risk TB groups in urban slums in India, the yield can be very small — only 0.8% of screened individuals were diagnosed with TB [9]. With an estimated 1 million undiagnosed TB cases in India, efficient active screening is the need of the hour [9].

Our *first contribution* is a model of the active screening problem which considers the underlying disease dynamics. We focus on recurrent infectious diseases with a latent stage (SEIS model of disease [26]), such as TB. Individuals can be susceptible (S) (currently healthy, but may become exposed), exposed (E), or infected (I). We consider diseases for which there is no means to achieve permanent immunity, either through vaccination or one time infection. As for TB, we assume treatment is effective for both exposed and infected individuals, making the individuals healthy (though again susceptible). Health workers are uncertain about the health state of individuals and have a small budget relative to population size for active screening. To the best of our knowledge, models of multi-round active screening for SEIS diseases are missing in the AI literature.

Our *second contribution* is a novel algorithm—**T**argeted **R**esolution of **A**ctive diseases using **C**ommunities and **E**igenvalues (TRACE)—to guide scalable active screening. In TRACE, we use network community structure to form a community graph, and then we select nodes to screen by maximizing the reduction of the largest eigenvalue of a variant of the community graph. TRACE takes into account the underlying disease dynamics and uncertainty of individuals' health states. TRACE is easily adaptable to most SEIS or SIS type diseases.

We illustrate the benefits of TRACE via extensive testing on real-world human contact networks against various baselines across a wide range of disease parameters (which also demonstrates its applicability to various other diseases).

2 DISEASE MODEL AND BACKGROUND

We first introduce the disease model notations for our problem. An individual can be in one of the following health states: S means that the individual is susceptible to disease (healthy), E means that the individual has been exposed and has latent disease, and I means that the individual is infected. We do not consider an explicit recovered or permanent immunity (R) state in our model, as this have been the focus of many prior studies. In diseases like Hepatitis A and measles, which follow a SEIR or SIR pattern, treated individuals may achieve permanent immunity by entering the recovered state [5, 24]. We focus on recurrent diseases, where permanent immunity is not possible (such as with TB, typhoid, and malaria), represented by SIS [1] or the more general SEIS [26] disease dynamics.

Disease Model: We adopt a SEIS model [26] for modeling the disease dynamics. TB and many other diseases follow a SEIS pattern, where treated individuals can relapse or become reinfected. The disease dynamics are therefore given by:

$$\text{Susceptible } (S) \xrightarrow{\alpha} \text{Exposed } (E)$$
$$\text{Exposed } (E) \xrightarrow{\beta} \text{Infected } (I)$$
$$\text{Infected } (I) \xrightarrow{c} \text{Susceptible } (S)$$

In the context of a graph of individuals, α is the edge-wise fixed probability of a susceptible (S) individual (node) being exposed (E) to the disease from an infected (I) neighbor, β is the fixed probability of an exposed (E) individual (node) becoming infected (I), and c is the probability of an infected (I) individual (node) voluntarily seeking and successfully completing treatment and returning to the susceptible S stage. We assume that the treatment takes place in one time period, where a period represents the duration needed for a complete treatment regimen (~half a year for TB).

Prior Approaches for Active Screening: Most previous work on active screening deals primarily with SIR or SEIR type diseases, often referred to as the *Vaccination Problem* [5, 24, 27, 30**?**], where permanent immunization (entry into R state) can be viewed as removing nodes from the graph [2, 20, 25]. Exploiting this idea, [20, 25] focus on immunization ahead of an epidemic and suggest a heuristic method of removing a set of k nodes based on the eigenvalues of the adjacency matrix. [30] considers the problem of selecting the best k nodes to immunize in a network after the disease has started to spread. These methods assume that the exact status of each node is known and deal with a single round of vaccination or screening. However, our paper focuses on multi-round active screening of SEIS diseases, where the complexity increases substantially due to lack of permanent immunity, existence of a latent stage, and uncertainty about the health states of all individuals. To the best of our knowledge, this complex setting has not been attempted previously in the AI literature. Generally, the problem of minimizing disease spread is different from the well-studied problem of influence maximization [**? ?**] as well, where one optimizes the selection of seeds or starting nodes for maximizing spread, as opposed to optimizing the selection of nodes on which to intervene in order to minimize spread.

3 ACTIVE SCREENING MODEL FORMULATION

Setup. We define k *active screening agents* that are to be deployed at every timestep t to diagnose and treat I and E individuals. Individuals are part of a contact network $G(V, E)$, and infection spreads via the edges in the network. There are $|V|$ individuals, and $N(i)$ denotes neighbors of individual i in the network. The network structure (graph) is known from the beginning ($t = 0$). Each individual (node) in the network is in one of the health states $\{S, E, I\}$. Let s_i^t denote the state of individual i at time t. In every round, the agents can either choose to screen a node i (action $a_i = 1$) or not ($a_i = 0$). Only k nodes can be screened in one round. A screened node is observed to be in state S, E, or I, and an unscreened node generates no observation. The agents maintain a belief about the state of every individual, starting with no information at $t = 0$. The beliefs about the health states evolve over time as the agents gain information about individuals (detailed later in this section).

Transition Dynamics. The probability of an individual undergoing a change in health state is given by:

$$T^0 = \begin{array}{c} \\ S \\ E \\ I \end{array} \begin{array}{c} \begin{array}{ccc} S & E & I \end{array} \\ \begin{bmatrix} q_j & 1-q_j & 0 \\ 0 & 1-\beta & \beta \\ c & 0 & 1-c \end{bmatrix} \end{array},$$

$$T^1 = \begin{array}{c} \\ S \\ E \\ I \end{array} \begin{array}{c} \begin{array}{ccc} S & E & I \end{array} \\ \begin{bmatrix} q_j & 1-q_j & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{array},$$

$$\text{and } q_j = (1-\alpha)^{|\{k \in N(j) \mid s_k^t = I\}|}$$

where, T^0 is the probability matrix for non-screened individuals and T^1 is the probability matrix for screened individuals. The rows denote the state at time t and the columns denote the state at $t + 1$. The transition probabilities follow the disease dynamics described earlier. In particular, q_j captures the probability that node j does not become exposed from his infected neighbors $\{k \in N(j) \mid s_k^t = I\}$. Both I and E individuals who are screened can be treated, but we assume E individuals do not seek treatment voluntarily since their disease is latent unlike I individuals who seek treatment voluntarily with the probability c. For model simplicity, we assume S individuals cannot directly transition directly to I state. This is not an extreme assumption for TB, where the overall duration with latent TB can be much longer than the round length (6 months).

Objective. Finally, we define a *reward* function $R(s^t) = \sum_j R(s_j^t)$, where $R(s_j^t)$ is defined as follows.

$$R(s_j^t) = \begin{cases} +1, & s_j^t = S \\ 0, & \text{otherwise} \end{cases}$$

The objective of the model is to choose the budget-limited actions at each time step in order to maximize the number of susceptible individuals over T time-steps: $\max \sum_{t=0}^{T} R(s^t)$, interpreted as the total number of disease-free half years [12]. This is closely related to another well known public health metric — QALY [**? ?**], where additionally a +1 reward is given to E individuals and a +0.66 reward to I individuals. We focus on maximizing health outcomes in this study and leave cost considerations to future work. An important part of the model is our belief updating approach, which is described next.

Belief States. We do not know the true health states of every individual at all times perfectly. We therefore model our belief of node i's health state as $\mathbf{b}_i^t = [b_{i,S}^t, b_{i,E}^t, b_{i,I}^t]$, where $b_{i,j}^t$ is the probability node i is in state j. This marginal representation of health state belief for each node i addresses scalability issues, as representations of the joint distribution of health state beliefs over all nodes can be prohibitively large. We assume marginal beliefs $\mathbf{b}_i^t$'s can be updated independently at each node. Such independence assumptions have been made in prior literature on the spread of contagion [8, 18] and experimentally found to have a minimal effect on outcomes.

Belief Update. We assume perfect observability of the health state s_i^t of any node when it is screened. We cannot observe the health state of a node at time t if we do not screen it at time t.

We update the belief for each individual (node) i who voluntarily come to the clinic to an intermediate belief state $\bar{\mathbf{b}}_i^t = [0, 0, 1]$. We also update the beliefs of actively screened individuals to an intermediate belief state $\bar{\mathbf{b}}_i^t \sim s_i^t$. We update the intermediate beliefs of the remaining individuals as:

$$\bar{\mathbf{b}}_i^t = \frac{[b_{i,S}^t, b_{i,E}^t, (1-c)b_{i,I}^t]}{b_{i,S}^t + b_{i,E}^t + (1-c)b_{i,I}^t}$$

For each node i that voluntarily came to a clinic or was actively screened, the final belief update is: $\mathbf{b}_i^{t+1} = [1, 0, 0]$ because the node will be successfully treated and returned to the susceptible state if it was in E or I state. For the remaining nodes, we update to $\mathbf{b}_i^{t+1}$ as follows:

$$\mathbf{b}_i^{t+1} = \bar{\mathbf{b}}_i^t \, \mathbf{\Gamma^t}, \text{ where}$$

$$\mathbf{\Gamma^t} = \begin{bmatrix} w_i^t & 1-w_i^t & 0 \\ 0 & 1-\beta & \beta \\ c & 0 & 1-c \end{bmatrix}, \; w_i^t = \prod_{j \in N(i)} (1 - \alpha \bar{b}_{j,I}^t).$$

This belief update procedure is an important and novel aspect of our proposed active screening model.

While the our model can be interpreted as a POMDP, it is slightly different from standard POMDP models, since in the active screening setting a screening action results in observing the current health states of the individual and not the individual's transitioned state. This difference can be handled straightforwardly, as in [4, 19] using a modified value iteration technique. However, we show in Section 6 that known POMDP approaches are not scalable for our problem.

4 MOTIVATION FOR TRACE

Given the problem setup, we motivate the need for the TRACE algorithm by showing that many prior approaches or simple extensions do not achieve the desired goal.

4.1 Eigenvalue Based Prior Approach

We first consider the circumstances under which diseases or epidemics die out on their own. In the absence of any intervention (action), the system is a discrete non-linear dynamical system. Such systems have been studied in prior work, and the following has been shown:

Proposition 1. *[18] Let λ_A^* denote the largest eigenvalue of the adjacency matrix A of the underlying graph, otherwise known as the spectral radius. Then, the epidemic dies out if and only if*

$$\frac{\alpha}{c} < \frac{1}{\lambda_A^*} \text{ and } \beta \neq 0 \; .$$

Remark: An observation is that the bound on λ_A^* above is same as derived for SIS model (without exposed E state) in earlier work [8]. This is because in the SEIS model, the E state must eventually become I if $\beta \neq 0$; thus, in the long run, E behaves similarly to I when $\beta \neq 0$ and there is no active intervention.

Permanent immunization can be viewed as removing nodes. Given the result above, one would wish to select the set of k nodes that reduces the largest eigenvalue the most. This is a NP-complete problem. [20, 25] suggest a heuristic that greedily removes k nodes one at a time, each time selecting the node that maximizes the reduction in the largest eigenvalue.

We also observe that the underlying problem is extremely hard to solve. In SIS networks, computing an individual's probability of infection and computing the expected number of infections are NP-hard [13, 21]. SIS is the relaxed version of the SEIS model, where $\beta = 1$. It is also known from [27] that given a network and limited resources, finding the optimal strategy for vaccinating a limited number of individuals (vaccination problem - SIR scenario), and quarantining a limited number of individuals (quarantining problem) are NP-hard. Also, given a network and limited resources, finding the optimal strategy for placement of a limited number of sensors for monitoring the course of an epidemic is NP-hard [21].

The Active Screening problem as defined in Section 3 is a generalized (harder) case of the above problems where we try to treat infected people without removing them from the graph since there is no permanent immunity and re-infection is possible (SEIS scenario). Based on Prop. 1, we also observe that a disease is unlikely to die out on its own in low-resource countries (c is low) with highly contagious diseases (high α), thus necessitating active screening.

4.2 Budgetary Threshold for Random Intervention

We can gain insight into how uncertainty in individuals' health states affects our problem by examining the fully-naive random screening strategy. We focus on the budget k, the number of nodes that can be screened and treated in one period. Intuitively, increasing k will lead to faster reduction of disease prevalence with random screening.

Lemma 1. *Assume that we know the infected patients belong to a set I_t in every round t such that $|I_t| \leq m$, where m is an arbitrary constant corresponding to the size of the network. Then, the epidemic dies out using k random interventions every round if $k > m(\lambda_A^* \alpha - c)$.*

Proof. The k random interventions among I_t nodes increase c by at least k/m and α is unchanged. Thus, the disease will die out if $\frac{\alpha}{c+k/m} < 1/\lambda_A^*$. □

Besides providing a threshold for k for which a naive intervention can achieve disease eradication, the above result can be understood as the price of limited information. Lower values of m, meaning more information (better estimate of the true health state), requires fewer random interventions to eradicate the disease. This underscores how uncertainty in the health states is an additional challenge when the number of interventions are limited.

4.3 Eigenvalue and Max Belief

Given the importance of information revealed above, a simple alternative to the eigenvalue approach could be to select k nodes with the top belief of being infected $b_{i,I}^t$ at every time step (denoted further as *Max Belief*). Unfortunately, both the eigenvalue method and Max Belief method have shortcomings in our dynamic problem. We demonstrate this through some observations for different classes of networks. In all the observations, $(\alpha, \beta, c) = (1, 1, 0)$. Also, for the sake of comparison, we assume all beliefs are close to the true states.

Observation 1. *There exists a class of graphs where the Max Belief method with a budget of $k \sim O(1)$ requires an expected $O(n!)$*

rounds to completely eradicate the disease whereas an eigenvalue-based method can eradicate the disease in an expected $O(n^2)$ rounds just with a budget of $k = 2$.

Justification. Consider a star graph (Figure 1a), where all the nodes are initially in I state. With a budget of 2, the eigenvalue method will choose the star center and one arbitrary node among non-central nodes to treat in every round. The disease will thus die out in an expected $\sum_{i=1}^{n-1} \frac{n-1}{i} \sim O(n^2)$ rounds. On the other hand, the Max Belief method will choose k nodes randomly among the nodes in state I. If the center node is not picked in every two rounds ($S \xrightarrow{1 \text{ round}} E \xrightarrow{1 \text{ round}} I$) before the disease dies out, the center will become infected, and after two more rounds the non-central nodes will be I except $2k$ nodes which can be either in S, E or I state (we ignore this w.l.o.g.). The probability of the center node being chosen every second round (because it takes two rounds to move from S to I state) is $\frac{k}{|I|}$ where $|I|$ is the total number of infected nodes in the round with the center being in I state. The probability of the center node being chosen every second round until the disease dies out is $\prod_{i=0}^{\frac{n}{2k-1}-1} \frac{k}{n-(2k-1)i}$. This gives the desired result.

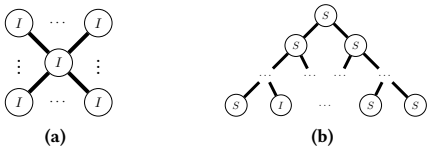


Figure 1: Comparing Eigenvalue and Max Belief

Observation 2. *There exists a class of graphs where an eigenvalue-based method can never eradicate the disease with a budget of $k < \frac{n}{2}$ whereas the Max Belief method can eradicate the disease in one round with a budget of $k \sim O(1)$.*

Justification. Consider a binary tree (Figure 1b), with $\Theta(k)$ leaf nodes in I state and others in S state. An eigenvalue-based method chooses the nodes that equally partitions the graph, and thus in this case it will start choosing from the root and go down the tree in breadth-first order, and reach the leaf nodes only after it has chosen all the $\frac{n-1}{2}$ parent nodes. Max Belief however can eradicate the disease in the first round by simply choosing k nodes which have the highest probability of being in I state, which are the infected leafs.

4.4 Community Based Approach

Infectious diseases such as TB are transmitted via close contact with an infected person, usually within communities [10]. Curing whole communities may potentially be an efficient way to reduce infection (can be interpreted as *graph shattering* [27]), since infection propagation is stopped for large sections of the graph. Also in our case, given the lack of additional information about the network like patient attributes, it is natural to utilize this approach. We also note that forming communities might enable us to reduce the largest eigenvalue, i.e. apply Algorithm **??**, in a scalable fashion.

However, we show in the following Observations that using communities alone can be both better or worse than Greedy or eigenvalue based approaches for different classes of graphs, further motivating the need for our algorithm, TRACE, which identifies communities in addition to considering beliefs and reducing the largest eigenvalue. The exact method of achieving scalability using communities is elucidated in the next section.

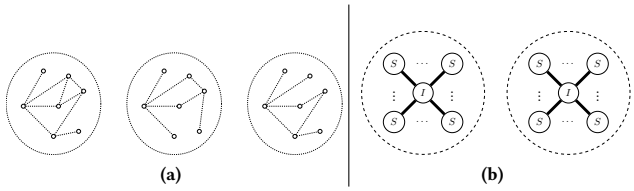


Figure 2: Comparing Eigenvalue, Community and Greedy

Observation 3. *There exists a class of graphs where an eigenvalue-based method can never eradicate the disease with a budget of k and the Greedy method requires an expected $O(|V|^k)$ rounds to completely eradicate the disease, but a community-based method can eradicate the disease in an expected $O(|V|^2)$ rounds.*

Justification. Let us consider a graph where there exists M disjoint clusters (Figure 2a), each of size less than or equal to the budget k with $k << M$, where M is the number of communities. All the nodes are in I state and are arranged in each cluster such that the the top k nodes, removal of which causes the most decrease in the largest eigenvalue, all lie in different clusters. In such graphs, it is evident that community-based algorithms can cure one community at a time and can achieve full eradication after an expected $M^2 \sim (|V|/k)^2 \sim O(|V|^2)$ rounds because a cured community cannot infect other communities. However, an eigenvalue-based technique may not choose communities as a whole and therefore, an eradication cannot be guaranteed unless the budget is increased to $|V|$ which is equal to the size of the graph. Similarly, the Greedy method may not choose communities as a whole and therefore takes an expected $\binom{|V|-1}{k-1}$ rounds to cure the first community, $\binom{|V|-k-1}{k-1}$ rounds to cure the second community, and so on, thus taking approximately $O(|V|^k)$ rounds to cure all the infected nodes.

Observation 4. *There exists a class of graphs where a community-based method can never eradicate the disease whereas the Greedy or eigenvalue-based method either can eradicate the disease in one round with a budget of k.*

Justification. Consider M disconnected star graphs (Figure 2b), where $M - 1$ stars are of size less than k and one star is of size k, and $k \leq M$. All the center nodes of the stars are in I state, and all the other nodes are in S state. With a budget of k, community-based algorithms will keep choosing the same star with k nodes thus never eradicating the disease. However, either the Greedy or eigenvalue-based method can directly choose the k center nodes in the first round and completely eradicate the disease in one shot.

5 TRACE ALGORITHM FOR ACTIVE SCREENING

We introduce a structured algorithm to generate an online POMDP policy—Targeted Resolution of Active diseases using Communities and Eigenvalues (TRACE)—that combines elements of the three approaches (Max Belief, and eigenvalue based, and community based methods) to identify the k individuals to actively screen at every time-step. The complete TRACE algorithm is shown in Algorithm 1. There are two distinct parts to this algorithm.

5.1 Community Formation and Intervention

As we do not know the true health state of all nodes in the network, we form communities using beliefs. The two step process is described below and is a part of Algorithm 1.

Node Type Estimation: We assign an attractiveness score to reflect the effectiveness of intervening on the node. If we knew the true health state of every node, then we would intervene only on the infected nodes as only these nodes spread infection. However, in the absence of such precise information, at every time-step the nodes are sorted according to a measure of possible benefit, defined as $R_i^t = \sigma b_{i,E}^t + b_{i,I}^t$ for each node i (line 2), where σ is an arbitrary parameter that controls the relative importance of E nodes relative to I nodes. The nodes with the highest one-third of R^t values are labeled g_1 (group 1), the next one-third to be g_2 (group 2), and the rest to be g_3 (group 3) (line 3).

Super-Node Creation: After labeling all nodes, locally similar nodes (nodes of the same label that share an edge) are clustered into a super-node iteratively. This process generates a set of super-nodes, each of which is labeled as g_1, g_2 or g_3 based on the labeling of its component nodes. There can be multiple super-nodes with the same label in the network. The $size_{\mathbf{u}}$ of a super-node $\mathbf{u}$ is the number of component nodes in the super-node. The weights of edges between nodes in different super-nodes are added to produce new inter-super-node edges. This super-nodes formation uses the known method of *graph coarsening* [11] (line 4). As an example, in Figure 3 we combine the two g_1, two g_2 and three g_3 nodes to form three super nodes with size two (and another with size one). These super-nodes emulate the communities of I, E and S in real-world networks. We refer to the resultant graph of super-nodes as the community graph, where the belief of each node $b_{\mathbf{u},S}^t$ is the average of $b_{v,S}^t$ of all component nodes v in super-node $\mathbf{u}$.

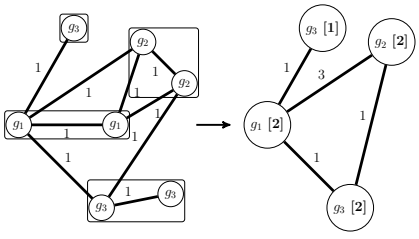


Figure 3: 4 super-nodes formed from 7 nodes

Next, we call the DynamicEigen sub-procedure to choose nodes to screen in the weighted community graph using *size* as weights on each super-node (line 5, where $\overline{\mathbf{A}}$ is the adjacency matrix of the community graph). The procedure returns a set of super-nodes

Algorithm 1 TRACE Algorithm

Input: Adjacency Matrix A of graph, Belief b^t, Budget k

1: **for** all $i \in \{1, \ldots, n\}$ **do**
2: $\quad R_i^t = \sigma b_{i,E}^t + b_{i,I}^t$
3: **Sort** R^t and label each node as g_1, g_2, or g_3
4: $\overline{\mathbf{A}}, \overline{\mathbf{b}}^t, size \leftarrow Coarsen(A, g_1, g_2, g_3, b^t)$
5: $\mathbf{U} \leftarrow$ DynamicEigen$(\overline{\mathbf{A}}, \overline{\mathbf{b}}^t, size, k)$
6: **if** $\sum_{\mathbf{u} \in \mathbf{U}} size_{\mathbf{u}} > k$ **then**
7: $\quad \mathbf{u}' \leftarrow$ the last selected super-node from $\mathbf{U}$
8: $\quad \kappa = k - \sum_{\mathbf{u} \in \mathbf{U} \setminus \mathbf{u}'} size_{\mathbf{u}}$
9: $\quad \underline{A}, \underline{b}^t \leftarrow$ remove all nodes in $\mathbf{U} \setminus \mathbf{u}'$ from A, b^t
10: $\quad a \leftarrow$ DynamicEigen$(\underline{A}, \underline{b}^t, \mathbf{1}, \kappa)$
11: Active screen nodes $\{v \mid v \in a$ or $v \in \mathbf{u}$ for $\mathbf{u} \in \mathbf{U} \setminus \mathbf{u}'\}$

where the total size (weight) is not lower than the budget k. If the total size is higher (line 6), we remove a super-node (line 7), compute left-over budget κ (line 8), modify the original graph by removing all nodes from the left-over super-nodes (line 9), and call the sub-procedure again to select κ nodes from the modified original graph with weights 1 on each node (line 10). It must be noted that our proposed DynamicEigen procedure is also one of the novel aspects of TRACE.

5.2 DynamicEigen Procedure

Next, we describe the DynamicEigen procedure, which is shown in Algorithm 2. Prior methods to minimize the largest eigenvalue greedily chose nodes to delete in order to generate a graph with lower maximal eigenvalue. Since we do not know which nodes are infected and can transmit infection with certainty, we augment this method by incorporating uncertainty. To motivate our approach, consider a *hypothetical scenario* where the state of each node is known for sure. We only wish to intervene on infected and exposed nodes, and S nodes do not effect neighboring nodes.

Using $A_{i,j} = A_{j,i} = 1$ to represent an edge from i to j in the adjacency matrix A of the input graph, we see that removing all edges from S nodes is same as multiplying the rows and columns of A corresponding to nodes in state S by zero. Then we can greedily choose among I and E nodes with the goal of reducing the largest eigenvalue of the adjacency matrix of the directed graph and return nodes that have total weights above the threshold k. While our intervention may be undone over time (treated nodes can be reinfected), repeated screenings may push the system towards lower disease prevalence.

Let us return to our problem setup, where we do not know the exact state of each node but rather have beliefs about each node. A natural extension of the hypothetical scenario above is to multiply the row of a node i in the adjacency matrix A by $1 - b_{i,S}^t$, the belief probability it is E or I (line 3). Algorithm 2 describes this approach. This is a softer version of making the row of all S nodes all zeros. Then, we perform greedy selection of nodes (lines 4-9) to reduce the largest eigenvalue of this matrix and to return nodes that have total weights above the threshold k.

Algorithm 2 DynamicEigen(A, b^t, w, k)

Input: Adjacency matrix A, belief b, function w for weight of each node, min total weight of nodes to remove k

$V \leftarrow$ Number of vertex of input graph
for all $i \in \{1, \ldots, V\}$ **do**
  $A_{i,:} = A_{i,:} * (1 - b_{i,S})$ ▷ Multiply i^{th} row
for all $i \in \{1, \ldots, V\}$ **do**
  $A' \leftarrow A$
  $\mathbf{A'_{i,:}} \leftarrow \mathbf{0}$, $\mathbf{A'_{:,i}} \leftarrow \mathbf{0}$ ▷ Remove i^{th} node
  $\lambda^i = LargestEigenvalue(A')$
Sort nodes $\langle v_1, \ldots, v_V \rangle$ corresponding to increasing λ^i
return first h nodes such that $\sum_{i=1}^{h} w(v_i) \geq k$

Now that we have combined community structure with belief states (denoted *Comm* in Section 6), we compare it to the DynamicEigen procedure (without super-node formation).

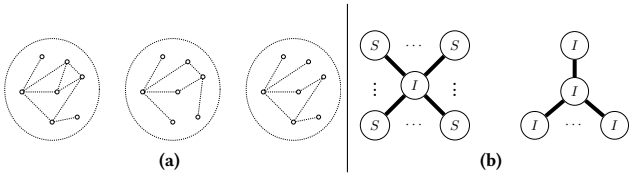


Figure 4: Comparing DynamicEigen and *Comm* approaches

Observation 5. *There exists a class of graphs where DynamicEigen without super-nodes can never completely eradicate the disease with a budget of k whereas the Comm algorithm can eradicate the disease in an expected $O(n)$ rounds.*

Justification. Consider a graph with M disjoint clusters (Figure 4a), each of size less than or equal to the budget k and $M > k$. All the nodes in all M communities are in I state. In such graphs, the *Comm* algorithm can treat one community at a time and achieve full eradication after $M \sim n/k \sim O(n)$ rounds as a community of S nodes cannot infect other communities. However, the DynamicEigen algorithm may not choose communities as a whole, therefore eradication cannot be guaranteed unless the budget is increased to n, which is equal to the size of the graph.

Observation 6. *There exists a class of graphs where the Comm algorithm with a budget of k requires an expected $O((n-n')!)$ rounds, to completely eradicate the disease whereas DynamicEigen without super-nodes can eradicate the disease in an expected $O(n')$ rounds with a budget of k, where n' is the size of the smaller star.*

Justification. Consider a graph with two stars of different sizes (Figure 4b) where the smaller star is of size $n' \geq k$ and the larger star has a size of $n - n'$. Initially, the center node in the larger star is in state I and the other nodes are in state S. All the nodes in the smaller star are in state I. The dynamic eigenvalue algorithm can eradicate the disease with just a budget of k in an expected $O(n')$ rounds by choosing both the stars' center and then choosing one non-central node and the center, or two non-central nodes in each round based on if the center node is in I state. However, the *Comm* algorithm will cluster the smaller star and cure all of them before choosing the I node in the larger star, where by then all of the nodes in the larger star would have been infected. Based on an analysis similar to Observation 1, we can conclude that the disease will die out in an expected $O((n - n')!)$ rounds.

Observation 7. *Suppose the belief states equal the actual health states and $(\alpha, \beta, c) = (1, 1, 0)$. Then, TRACE is guaranteed to perform better than or at least as well as its individual components, in terms of both budget and time, in all the classes of graphs discussed in the Observations.*

Proof. For example, in Figure 1a, in case of exact beliefs, it is guaranteed that TRACE will choose the central node since that is the best choice by eigenvalue (all I nodes have equal belief of [0,0,1]) and thereby eradicate the disease in $O(|V|^2)$ rounds with a budget of $k = 2$. Similarly, in Figure 1b, TRACE is guaranteed to choose all the k infected nodes since all the other nodes have zero belief of being in I state, thus eradicating the disease in one round. Thus, following Algorithm 1, we can similarly show that TRACE will in fact perform at least as well as its individual components in all the discussed classes of graphs (variants of trees, stars, and clusters). We omit the details for brevity. □

Thus, TRACE is able to leverage the advantages of each approach. Although these special graphs do not by themselves represent real-world human contact graphs, real graphs are formed from a combination of these special graphs. Estimating that the belief space representation is a reasonably accurate embedding of the information we do have (there is no misinformation in observations while screening), we hypothesize that TRACE's superior performance in these skeleton graphs can be extended to interpret good performance in realistic graphs as well. This hypothesis is validated via experiments.

6 EXPERIMENTS

We consider three real-world datasets on which we perform experiments.

(1) **India** network [6]: A human contact network with $n = 202$ nodes, collected from a rural village in India, a setting in which TB active screening may take place ($1/\lambda_A^* \sim 0.095$).
(2) Infectious **Exhibition** network [14]: A real-world human contact network with $n = 410$ nodes, collected during an artificial simulation of contagion and containment at an exhibition ($1/\lambda_A^* \sim 0.043$).
(3) **Irvine** network [15]: An online social network with $n = 1899$ nodes, constructed from sent messages between the users of an online community of students from UC Irvine ($1/\lambda_A^* \sim 0.021$).

As discussed in Section 3, we first attempt to solve our special POMDP using the state-of-the-art modified POMCP algorithm [19]. We show in Figure 5 that POMCP takes exponential time with increasing n and fails to scale up beyond 10 nodes (India network) for fixed values of k and T while TRACE is able to generate an online POMDP policy for the whole network without exponential increase in runtime. Factored POMDPs [29] and newer algorithms

like DESPOT [22] also fail to scale up beyond a few nodes due to memory overflow. All results are averages over 20 simulation runs.

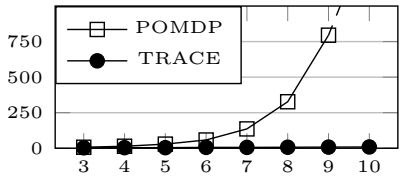


Figure 5: Runtime (s) v/s Number of nodes (n); $k = 3, T = 10$

Settings. Next, we analyze TRACE's performance under various α, β, c settings. α, β, c may depend on social contact patterns and biological factors which may vary across populations [23]. We explore a range of these parameters to show disease behavior under a variety of scenarios. Since eradication does not depend on β (by Proposition 1), we vary only α, c and fix $\beta = 0.25$ for the experiments. The passive treatment rate c may vary widely, as it depends on resource availability (clinic accessibility, outreach campaigns, etc.). In all simulations, the budget is $k = 5\%$ of the total population, and $\sigma = 0.5$.

Setup. In the real world, active screening is performed only after conducting initial surveys on the prevalence and incidence of the disease. To simulate this, we run our experiments in two stages.

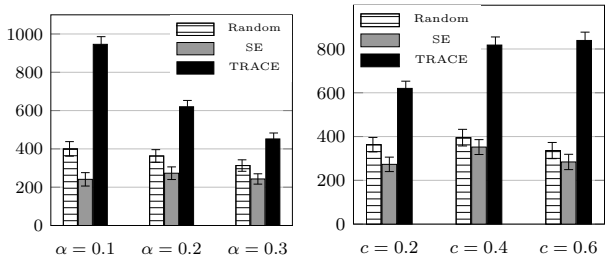
(1) Stage 1 (**Survey Stage**), starts at $t = 0$ with equal number of S, E, I individuals and ends at $t = 10$. No active screening is done and the disease evolves naturally. The initial belief b^0 for all nodes is assumed to be $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ since we have no prior information. Beliefs are updated when individuals come to the clinic voluntarily (with probability c).
(2) Stage 2 (**Active Screening Stage**), we consider various screening algorithms. We perform active screening from $t = 11$ to $t = 30$ to represent 10 years of time (each round is 6 months [7]). We compare the benefit of these screening strategies over and above no intervention (**None**), where in **None** the evolution of the health states is based on disease dynamics with no active screening.

Comparison with baselines. Given the lack of previous algorithms, Figures 6 and 7 show the performance of TRACE against simple baselines:

(1a) **Random**: Randomly select nodes for active screening.
(1b) **Static Eigen (SE)**: Choose the nodes using Algorithm 2 after removing lines 2 & 3 (no belief information), on the network (no super-node formation). This baseline uses only the graph structure information.

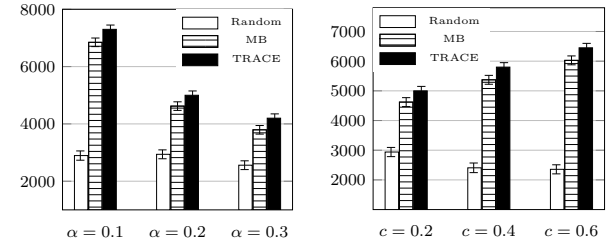
TRACE provides significant improvement over None compared to SE and Random ($p < 0.05$). The improvement is also practically significant (Cohen's $d > 1$: large effect).

Comparison with individual components. Figure 8 shows the performance of the three approaches that were combined to form TRACE, illustrating that no single approach is solely responsible for TRACE's performance. We compare the increase in $\sum_{t=0}^{t=30} |S|_t$ for each approach over None. TRACE's performance is both statistically and practically significant ($p < 0.05$ and Cohen's $d \sim 0.6$: medium effect) when compared to the three approaches:



(a) Varying α ($c = 0.2$) **(b) Varying c ($\alpha = 0.2$)**

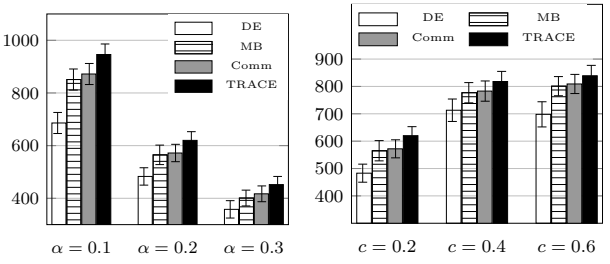
Figure 6: Increase in $\sum_{t=0}^{t=30} |S|_t$ for naive baselines and TRACE over None (India network)



(a) Varying α ($c = 0.2$) **(b) Varying c ($\alpha = 0.2$)**

Figure 7: Increase in $\sum_{t=0}^{t=30} |S|_t$ for naive baselines and TRACE over None (Irvine network)

(2a) **Dynamic Eigen (DE)**: Choose the nodes using just Algorithm 2 without any super-node formation.
(2b) **Max Belief (MB)**: Choose the nodes with the higher *belief* of being infected in that time-step, i.e. $b^t_{i,I}$.
(2c) **Community (Comm)**: Choose the nodes by a 0-1 knapsack algorithm (knapsack weight = budget k) after super-node formation.



(a) Varying α ($c = 0.2$) **(b) Varying c ($\alpha = 0.2$)**

Figure 8: Performance by TRACE component (India network)

Further, we analyze the minimum additional budget required to achieve performance comparable to TRACE in Figure 9, revealing the budgetary savings from using TRACE. TRACE with all its components produces significant savings over attempting to use each component alone.

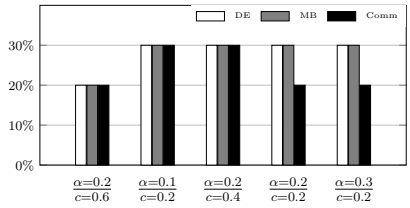
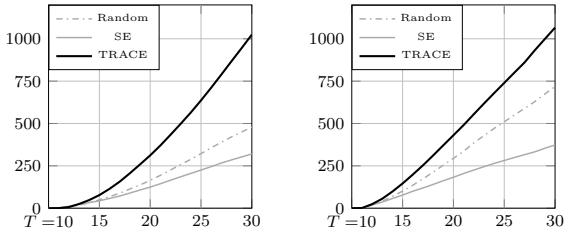


Figure 9: Minimum extra budget (in %) required to match performance of TRACE (India network)

The synergy of belief states, eigenvalues and community gives TRACE a clear advantage on both the datasets (Figure 10), where we see an increasing divergence over time in the performance of TRACE compared to **Random** and **SE**.



(a) India network **(b) Exhibition network**

Figure 10: Increase in $\sum_{t=0}^{t=T} |S|_t$ over None for varying T ($\alpha = 0.1, \beta = 0.25, c = 0.2$)

7 CONCLUSION

We proposed a novel active screening model and an algorithm (TRACE) to facilitate multi-round active screening for recurrent diseases. Unlike existing works in AI literature, the Active Screening model incorporates uncertainty of health states as well as the SEIS disease complexities of no permanent cure and a latent stage. TRACE performs significantly better, in a scalable fashion, than the baselines and each of its components individually in a variety of real-world inspired settings.

Future directions include incorporating more complex disease models (e.g. including maternal immunity, carrier states etc.), including birth and death processes, and introducing patient heterogeneity (age, gender, medical history and other features) and costs of treatment and screening into the model.

REFERENCES

[1] Benjamin Armbruster and Margaret L Brandeau. 2007. Optimal mix of screening and contact tracing for endemic diseases. *Mathematical biosciences* 209, 2 (2007), 386–402.

[2] James Aspnes, Kevin Chang, and Aleksandr Yampolskiy. 2006. Inoculation strategies for victims of viruses and the sum-of-squares partition problem. *J. Comput. System Sci.* 72, 6 (2006), 1077–1093.

[3] National Tuberculosis Controllers Association et al. 2005. Guidelines for the investigation of contacts of persons with infectious tuberculosis. Recommendations from the National Tuberculosis Controllers Association and CDC. *MMWR. Recommendations and reports: Morbidity and mortality weekly report. Recommendations and reports* 54, RR-15 (2005), 1.

[4] Turgay Ayer, Oguzhan Alagoz, and Natasha K Stout. 2012. OR ForumâĂŤA POMDP approach to personalize mammography screening decisions. *Operations Research* 60, 5 (2012), 1019–1034.

[5] Frank G Ball, Edward S Knock, and Philip D OâĂŹNeill. 2015. Stochastic epidemic models featuring contact tracing with delays. *Mathematical biosciences* 266 (2015), 23–35.

[6] Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. 2013. The diffusion of microfinance. *Science* 341, 6144 (2013), 1236498.

[7] CDC. 2011. Tuberculosis: General Information. (2011). https://www.cdc.gov/tb/publications/factsheets/general/tb.pdf

[8] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. 2008. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)* 10, 4 (2008), 1.

[9] Palanivel Chinnakali, Pruthu Thekkur, Gomathi Ramaswamy, Kalaiselvi Selvaraj, et al. 2016. Active screening for tuberculosis among slum dwellers in selected urban slums of Puducherry, South India. *Annals of Tropical Medicine and Public Health* 9, 4 (2016), 295.

[10] Collette N Classen, Robin Warren, Madeleine Richardson, John H Hauman, Robert P Gie, James HP Ellis, Paul D van Helden, and Nulda Beyers. 1999. Impact of social interactions in the community on the transmission of tuberculosis in a high incidence area. *Thorax* 54, 2 (1999), 136–140.

[11] Bruce Hendrickson and Robert W Leland. 1995. A Multi-Level Algorithm For Partitioning Graphs. *SC* 95, 28 (1995).

[12] IHME. 2010. *The Global Burden of Disease: Generating Evidence, Guiding, Policy.* World Bank.

[13] Jing Kjeldsen. 2013. *The probability for infection in SIR and SIS networks.* Master's thesis.

[14] KONECT. 2017. Infectious network dataset – KONECT. http://konect.uni-koblenz.de/networks/sociopatterns-infectious

[15] KONECT. 2017. Uc irvine messages network dataset – KONECT. http://konect.uni-koblenz.de/networks/opsahl-ucsocial

[16] K Kranzer, H Afnan-Holmes, K Tomlin, Jonathan E Golub, AE Shapiro, A Schaap, EL Corbett, K Lönnroth, and JR Glynn. 2013. The benefits to communities and individuals of screening for active tuberculosis disease: a systematic review [State of the art series. Case finding/screening. Number 2 in the series]. *The international journal of tuberculosis and lung disease* 17, 4 (2013), 432–446.

[17] E Mitchell, Saskia den Boon, and K Lonnroth. 2013. Acceptability of household and community-based TB screening in high burden communities: a systematic literature review. WHO.

[18] B Aditya Prakash, Deepayan Chakrabarti, Nicholas C Valler, Michalis Faloutsos, and Christos Faloutsos. 2012. Threshold conditions for arbitrary cascade models on arbitrary networks. *Knowledge and information systems* 33, 3 (2012), 549–575.

[19] Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. 2016. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 123–131.

[20] Sudip Saha, Abhijin Adiga, B Aditya Prakash, and Anil Kumar S Vullikanti. 2015. Approximation algorithms for reducing the spectral radius to control epidemic spread. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 568–576.

[21] Michael Shapiro and Edgar Delgado-Eckert. 2012. Finding the probability of infection in an SIR network is NP-Hard. *Mathematical biosciences* 240, 2 (2012), 77–84.

[22] Adhiraj Somani, Nan Ye, David Hsu, and Wee Sun Lee. 2013. DESPOT: Online POMDP planning with regularization. In *Advances in neural information processing systems*. 1772–1780.

[23] Sze-chuan Suen, Eran Bendavid, and Jeremy D Goldhaber-Fiebert. 2014. Disease control implications of India's changing multi-drug resistant tuberculosis epidemic. *PloS one* 9, 3 (2014), e89822.

[24] Chengjun Sun and Ying-Hen Hsieh. 2010. Global analysis of an SEIR model with varying population size and vaccination. *Applied Mathematical Modelling* 34, 10 (2010), 2685–2697.

[25] Hanghang Tong, B Aditya Prakash, Tina Eliassi-Rad, Michalis Faloutsos, and Christos Faloutsos. 2012. Gelling, and melting, large graphs by edge manipulation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 245–254.

[26] P Van den Driessche, M Li, and J Muldowney. 1999. Global stability of SEIRS models in epidemiology. *Canadian Applied Mathematics Quarterly* 7 (1999), 409–425.

[27] Nan Wang. 2005. *Modeling and analysis of massive social networks.* Ph.D. Dissertation.

[28] WHO. 2017. Global tuberculosis report 2017. (2017).

[29] Jason D Williams, Pascal Poupart, and Steve Young. 2005. Factored partially observable Markov decision processes for dialogue management. In *Proc. IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. 76–82.

[30] Yao Zhang and B Aditya Prakash. 2015. Data-aware vaccine allocation over large networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10, 2 (2015), 20.

Validation of Network-Dependent Epidemic Processes

A Study of Dr. Snow's Seminal Cholera Dataset*

Philip E. Paré
University of Illinois
Urbana, Illinois 61801
philpare@illinois.edu

Ji Liu
Stony Brook University
Stony Brook, New York 11794
ji.liu@stonybrook.edu

Carolyn L. Beck
University of Illinois
Urbana, Illinois 61801
beck3@illinois.edu

Tamer Başar
University of Illinois
Urbana, Illinois 61801
basar1@illinois.edu

Angelia Nedić
Arizona State University
Tempe, Arizona 85281
angelia.nedich@asu.edu

ABSTRACT

Models of spread processes over non-trivial networks are commonly motivated by modeling and analysis of biological networks, computer networks, and human contact networks. However, identification of such models has not yet been explored in detail, and the models have not been validated by real data. In this paper, we present a sufficient condition for asymptotic stability of the healthy equilibrium, show that the condition is necessary and sufficient for uniqueness of the healthy equilibrium, and present a result on learning the ratio of the spread parameters. Finally, we employ John Snow's seminal work on cholera epidemics in London in the 1850's to validate an approximation of a well-studied network-dependent susceptible-infected-susceptible (SIS) model.

KEYWORDS

SIS epidemic processes, model validation, data-driven analysis

ACM Reference format:
Philip E. Paré, Ji Liu, Carolyn L. Beck, Tamer Başar, and Angelia Nedić. 1997. Validation of Network-Dependent Epidemic Processes. In *Proceedings of ACM KDD epiDAMIK'18 - International Workshop on Epidemiology meets Data Mining and Knowledge Discovery, London, UK, August 2018 (epiDAMIK'18),* 4 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

Mathematical models of virus spread have been studied for centuries [2]. Recently these models have been extended to include network structure. In this work we focus on SIS models with infection parameters β_i and a healing rates δ_i. A virus model is called *homogeneous* if the infection and healing rates are the same for every agent, and *heterogeneous* if they are different for each agent. In this work, we focus on discrete-time SIS models, mainly for the more general, heterogeneous models. For reviews on epidemic processes see [8, 10].

*The full version of this work is available in [9].

epiDAMIK'18, London, UK
 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

While parameter estimation of epidemic spread with real data has been carried out for some models [6, 7, 15], the previous work has either not had network structure included or employed a large probabilistic model. Ignoring network structure is tantamount to making a strong simplifying assumption, and using a full probabilistic model can become very computationally expensive as the size of the network grows. For these reasons we focus on a nonlinear network-dependent ordinary differential equation model. To the best of our knowledge, no work has been done on the identification of spread parameters from data for these models. Many virus spread papers using these models have claimed to use real data to test their models, but no true validation of non-trivial network-dependent SIS spread models has been done. Previous work has used real data to identify underlying network structure, however there have been no prior efforts that have considered spread process data and identification over these networks. [4, 14].

We use the cholera dataset compiled by John Snow in [12] to validate the spread model analyzed in this work. Dr. Snow mapped the deaths caused by cholera in the Soho District of London in 1854 to illustrate that the infection was being spread by contaminated water via a specific pump, the Broad Street pump, and not via the air, as was the belief at the time. This seminal work by Snow has led to the modern day field of epidemiology [3]. While now, partially due to Snow, we understand cholera, how it spreads, and how to mitigate it, this illness is still a serious problem in poorer parts of the world today, highlighted by the current outbreak in Yemen where there have been over one million suspected cases of cholera and over 2,270 cholera-related deaths since the end of April 2017 [1].

John Snow's original spatial dataset of the cholera epidemic is static and does not contain time series data. Shiode *et al.* created spatial time series data, presented in [11] using additional sources and some statistical methods. However, Shiode *et al.* did not perform any dynamic analysis on their dataset, and have not made the dataset publicly available. We use a technique developed in the analysis section herein, combined with several strong but reasonable assumptions, to reproduce time series data, and in so doing, validate the model with the dataset. As far as we know, this is the first attempt to study Snow's cholera dataset from a dynamical systems' perspective to validate models of epidemic processes.

1 SIS MODEL

We focus on a discrete-time SIS model. The state x_i can correspond to the probability of infection of the ith agent [13] or the infected

proportion of group i [5]. For the identification of the spread process parameters in Section 3 we employ the latter case. We model the system dynamics by

$$x_i^{k+1} = x_i^k + h\left((1 - x_i^k)\beta_i \sum_{j=1}^{n} a_{ij} x_j^k - \delta_i x_i^k\right), \tag{1}$$

where k is the time index and $h > 0$ is the sampling parameter. We write (1) in matrix form as

$$x^{k+1} = x^k + h((I - X^k)BA - D)x^k, \tag{2}$$

where $X^k = diag(x^k)$, $B = diag(\beta_i)$, and $D = diag(\delta_i)$. Note that A is the matrix of a_{ij}'s and is not necessarily symmetric.

For the model to be well-defined we make several assumptions.

Assumption 1. *For all $i \in [n]$, we have $x_i^0 \in [0, 1]$.*

Assumption 2. *For all $i \in [n]$, we have $\beta_i \geq 0$, $\delta_i \geq 0$ and, for all $j \in [n]$, $a_{ij} \geq 0$.*

Assumption 3. *For all $i \in [n]$, $h\delta_i \leq 1$ and $h\beta_i \sum_{j \neq i} a_{ij} \leq 1$.*

Lemma 1.1. *For the system in* (2)*, under the conditions of Assumptions 1, 2, and 3, $x_i^k \in [0, 1]$ for all $i \in [n]$ and $k \geq 0$.*

Lemma 1.1 implies that the set $[0, 1]^n$ is positively invariant with respect to the system defined by (2). Since x_i denotes the fraction of group i infected, or is an approximation of the probability of infection of individual i and $1-x_i$ denotes the fraction of group i that is healthy, or is an approximation of the probability of individual i being healthy, it is natural to assume that their initial values are in the interval $[0, 1]$, since otherwise the values will lack any physical meaning for the epidemic model considered here. Therefore, we focus on the analysis of (2) only on the domain $[0, 1]^n$.

We also make the following assumption to ensure *non-trivial* virus spread.

Assumption 4. *We have $h \neq 0$ and $\exists i \neq j$ s.t. $\beta_{ij} > 0$.*

Note that we do not assume the healing rates to be nonzero. This allows for the possibility of SI (susceptible-infected) models.

2 ANALYSIS

For analysis purposes we need an assumption on the structure of the BA matrix. A square matrix is called *irreducible* if it cannot be permuted to a block upper triangular matrix.

Assumption 5. *The matrix BA is irreducible.*

Note that this assumption is equivalent to the underlying graph being strongly connected.

Theorem 2.1. *Suppose that Assumptions 1-5 hold for* (2)*. If $\rho(I - hD + hBA) \leq 1$, then the healthy state is asymptotically stable with domain of attraction $[0, 1]^n$.*

Proposition 1. *Suppose that Assumptions 1-5 hold. If $\rho(I - hD + hBA) > 1$, then* (2) *has two equilibria, $\mathbf{0}$ and x^*, where $x^* \gg \mathbf{0}$.*

Theorem 2.2. *Under Assumptions 1-5, the healthy state is the unique equilibrium of* (2) *if and only if $\rho(I - hD + hBA) \leq 1$.*

The following corollary shows that the ratio of the spread parameters can be recovered for the heterogeneous case with different δ_i's and β_i's for each agent (and includes the homogeneous case as a special case) if A and the endemic state are known.

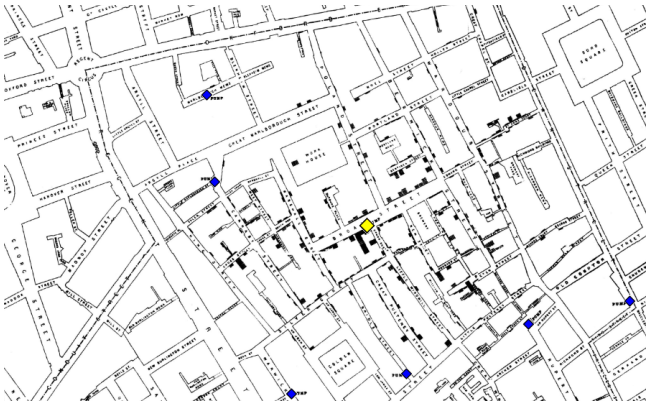


Figure 1: Map of cholera spread in London in 1854 compiled by John Snow [12]: healthy water pumps, the contaminated pump, and household deaths are depicted by blue diamonds, the yellow diamond, and black rectangles, respectively.

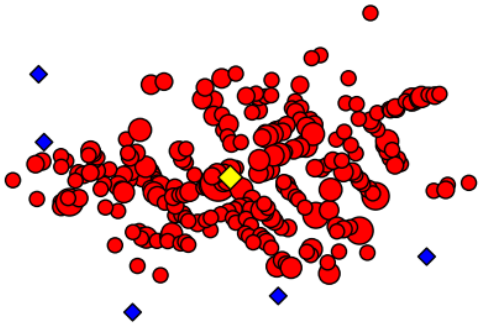


Figure 2: Digitization of Figure 1: The healthy water pumps, the contaminated pump, and the deaths are depicted by blue diamonds, the yellow diamond, and red dots with the diameters scaled by the number of deaths, respectively.

Corollary 2.3. *Considering the model in* (1) *under Assumptions 1-5, if A and the endemic state, $x^* \gg 0$, are known, then*

$$\frac{\delta_i}{\beta_i} = \frac{(1 - x_i^*)}{x_i^*} \sum_{j=1}^{n} a_{ij} x_j^*. \tag{3}$$

We will use the above corollary in the validation work that follows.

3 VALIDATION: SNOW DATASET

Now we employ the seminal cholera dataset collected by John Snow [12] for validation of the model in (1).

3.1 Snow Dataset

Snow depicted the number of deaths per household caused by cholera in the Soho District of London in 1854 on a map of the area. In Figure 1, the original map is shown, where each small rectangle corresponds to one death at that address. Snow created this map to illustrate to officials that the cholera epidemic was being spread by infected water from the Broad Street pump (the yellow diamond), and not through the air, the common belief at that time. We have plotted this data in Figure 2, with diamonds indicating the water

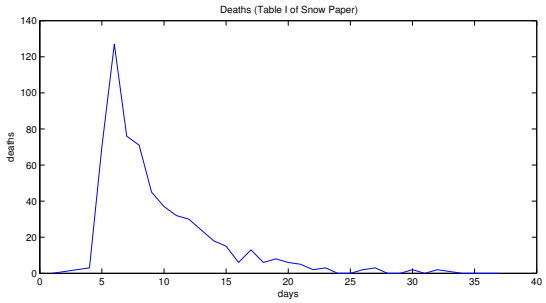
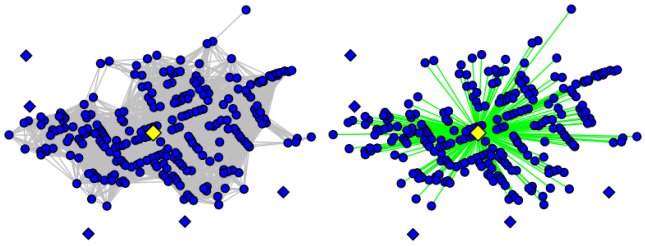


Figure 3: Total deaths per day in the Soho District of London in 1854, compiled by John Snow (from Table I in [12]).



(a) $A^{(1)}$ **from** (5) **(b)** $A^{(2)}$ **from** (6)

Figure 4: Initial condition of simulations with graph structures: blue circles indicate healthy households and the yellow diamond indicates the infected pump.

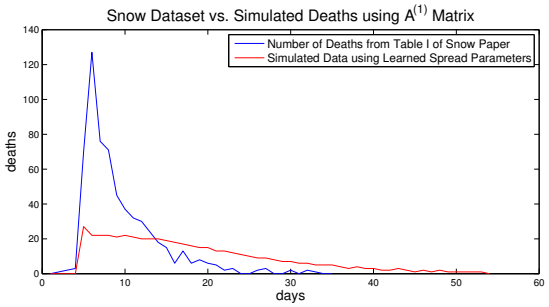


Figure 5: Comparison of Figures 3 and the simulated data using the learned parameters from the data in Figure 2, employing Corollary 2.3 and $A^{(1)}$ from (5): Note that the model does not capture the behavior of the system. The Euclidean distance between the two plots is 146.52, and the infinity norm is 105.

pumps and red dots indicating deaths. The dataset is comprised of 250 households with at least one death. Snow also documented the cumulative deaths per day in Table I of [12], plotted in Figure 3. The time of deaths for each address is not recorded. The total cumulative deaths in the table is 616, but the total number of deaths on the map are 489. Therefore, there is a discrepancy of 127 deaths, whose household addresses are not included in the map.

3.2 Spread Validation

For the validation, each household with a death recorded by Snow in the map in Figure 1 corresponds to a node in the model. The last node in the model corresponds to the contaminated pump, the one on Broad Street, and we do not include the healthy water pumps in the model. We realize that ignoring the households with no recorded deaths and ignoring the healthy pumps are nontrivial assumptions. However, as was noted by Snow, many residents fled the city once they became aware of the outbreak [12]. For the households that did not flee, we assume they either had such a high healing rate that their inclusion would have been trivial and/or that these households exclusively drank from another pump and did not closely associate with neighbors who did drink from the Broad Street pump. Despite these (and subsequent) relatively strong assumptions, the validation results are quite promising.

The state of the system, x^k, is the percentage of total deaths in each household up to time k. The epidemic equilibrium of the system, which we call x^*, was calculated from the data in Figure 2, for the first attempt, by dividing the total number of deaths in each household by 20, and therefore assuming that each household has 20 members. This number was chosen because the maximum number of deaths was 15. For the last attempt we approximated the household sizes using Figure 1 in [11]; see Table 1. The last element of x^*, corresponding to the contaminated pump, was set to $\frac{19}{20}$.

We employed Corollary 2.3 to calculate the $\frac{\delta_i}{\beta_i}$ values. Then for simulation we set $\beta_i = 1$ for all i and chose h as large as possible while still meeting Assumption 3. For the initial condition in the simulations, we began with the Broad Street pump infected and all the households healthy:

$$x^0 = \begin{bmatrix} 0 & \dots & 0 & 1 \end{bmatrix}^\top. \tag{4}$$

This initial condition is shown in Figure 4 (as well as the two considered graph structures), where the contaminated pump is depicted as a yellow diamond. As a consequence of these assumptions, our tuning parameter for adjusting the learned δ_i parameters, and consequently the spread behavior, was the connectivity matrix A.

For the first attempt, we designed $A^{(1)}$ such that

$$a_{ij}^{(1)} = \begin{cases} 1, & \text{if } \|z_i - z_j\| < r, \\ 1, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where z_i is the location of household i and r was smallest number such that the graph was connected (shown in Figure 4a). Using the $\frac{\delta_i}{\beta_i}$ values derived using $A^{(1)}$, we simulated the system, using (1). To meet the constraints of Assumption 3, we had to set $h = \frac{1}{175}$. To create a plot of deaths per day, we multiplied the state of the system, i.e., the percentage of deaths in each household up to that point, by the household sizes (assumed to be 20), rounded to the nearest integer, took the difference between the states of each time step (since the state represents cumulative number of deaths up to that point), and then summed every three time series points (due to the small h value), therefore assuming that each time series point corresponds to a third of a day. Note that this approach does not capture the behavior of the system very well as it is very different than the dataset, as depicted in Figure 5. The Euclidean distance between the two plots is 146.52, and the infinity norm is 105.

Household Sizes	
Range in [11]	Estimate
0-4	4
5-9	7
10-14	12
15-24	20
24-403	25*

Table 1: Estimates for household sizes from Figure 1 in [11] used in the simulation with $A^{(2)}$: *The workhouse population was set to 403.

For the final attempt we changed to heterogeneous household sizes, using Figure 1 in [11] to approximate these values. We removed all edges except the self loops and the binary directed edges from the pump to every household with at least one death. The connection from the pump to the workhouse was set to $\frac{1}{10}$ because they had their own well and only a small fraction of the 403 residents drank from the Broad Street pump [12]. Therefore

$$A^{(2)} = \begin{bmatrix} 1 & 0 & \dots & 0 & 1 \\ 0 & 1 & \dots & 0 & \vdots \\ 0 & 0 & \ddots & 0 & \frac{1}{10} \\ 0 & 0 & \dots & 1 & \vdots \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}. \quad (6)$$

We found via simulation that as long as the edge weight corresponding to the workhouse was less than or equal to 0.45 then the results were very similar.

Plotting the data from Figure 3 and the simulated data using the learned parameters from the data in Figure 2, employing Corollary 2.3 and $A^{(2)}$ from (6) on the same plot for comparison in Figure 6 shows that we capture the behavior of the outbreak quite well. The Euclidean distance between the two plots is 75.16, and the infinity norm is 70. One of the reasons for this discrepancy is due to the fact that we used the spatial dataset in Figures 1-2, which had only 489 documented deaths, while the cumulative data from Table I in [12], shown in Figure 3 and the blue line in Figure 6, has a total of 616 deaths. The difference of 127 has caused the discrepancy. The lack of the address information for the additional 127 deaths is one of the reasons the plots are not identical. However, the discrepancy is distributed fairly evenly across the whole sample time. Consequently, we have shown that the model in (1) captures the behavior of the cholera epidemic from John Snow's 1854 dataset very well. Note that the fact that $A^{(2)}$ from (6) performs the best supports Snow's hypotheses that the Broad Street pump was the source of the cholera outbreak, and that cholera does not spread easily between people or the air, which is known to be true today.

4 CONCLUSION

We have provided necessary and sufficient conditions for uniqueness of the healthy equilibrium, conditions for the existence of an endemic state., and a necessary condition for asymptotic stability of the healthy state. We use a corollary of this analysis to recover the ratio of the virus spread parameters. Using this corollary we have validated a discrete-time, network-dependent SIS virus spread model using John Snow's seminal cholera dataset with very good results. In future work, would like to find other datasets to help further validate the SIS models. We would like to further study identification of the spread model accounting for noise in the data.

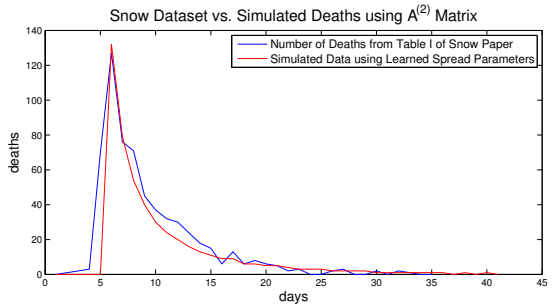


Figure 6: Comparison of Figure 3 and the simulated data using the learned parameters from the data derived using $A^{(2)}$ in (6): Note that there is a difference in the magnitude, but the general shapes are very similar.

REFERENCES

[1] Reema Al Yusfi, Malika Bouhenia, and Lauren O'Connor. 2018. Weekly Epidemiological Bulletin W14 2018. (2018). http://www.emro.who.int/images/stories/yemen/week_14.pdf?ua=1.

[2] Daniel Bernoulli. 1760. Essai dfiune nouvelle analyse de la mortalité causée par la petite vérole et des avantages de lfiinoculation pour la prévenir. *Histoire de lfiAcad. Roy. Sci.(Paris) avec Mém. des Math. et Phys. and Mém* (1760), 1–45.

[3] Ruth Bonita, Robert Beaglehole, and Tord Kjellström. 2006. *Basic epidemiology*. World Health Organization.

[4] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. 2008. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)* 10, 4 (2008), 1.

[5] A Fall, Abderrahman Iggidr, Gauthier Sallet, Jean-Jules Tewa, and others. 2007. Epidemiological models and Lyapunov functions. *Math. Model. Nat. Phenom* 2, 1 (2007), 62–68.

[6] Matt J Keeling, , and *et al.* 2001. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* 294, 5543 (2001), 813–817.

[7] Hongyu Miao, Xiaohua Xia, Alan S Perelson, and Hulin Wu. 2011. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM review* 53, 1 (2011), 3–39.

[8] Cameron Nowzari, Victor M Preciado, and George J Pappas. 2016. Analysis and Control of Epidemics: A Survey of Spreading Processes on Complex Networks. *IEEE Control Systems Magazine* 36, 1 (2016), 26–46.

[9] P. E. Paré, J. Liu, C. L. Beck, B. E. Kirwan, , and T. Başar. 2018. Discrete Time Virus Spread Processes: Analysis, Identification, and Validation. conditionally accepted to *IEEE Transactions on Control Systems Technology: System Identification and Control in Biomedical Applications* (2018). arXiv:1710.11149 [math.OC].

[10] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. 2015. Epidemic processes in complex networks. *Reviews of modern physics* 87, 3 (2015), 925.

[11] Narushige Shiode, Shino Shiode, Elodie Rod-Thatcher, Sanjay Rana, and Peter Vinten-Johansen. 2015. The mortality rates and the space-time patterns of John Snow's cholera epidemic map. *International journal of health geographics* 14, 1 (2015), 21.

[12] John Snow. 1855. *On the mode of communication of cholera.* John Churchill.

[13] Piet Van Mieghem, Jasmina Omic, and Robert Kooij. 2009. Virus spread in networks. *IEEE/ACM Transactions on Networking* 17, 1 (2009), 1–14.

[14] Yan Wan, Sandip Roy, and Ali Saberi. 2008. Designing spatially heterogeneous strategies for control of virus spread. *IET Systems Biology* 2, 4 (2008), 184–201.

[15] Ling Xue, H Morgan Scott, Lee W Cohnstaedt, and Caterina Scoglio. 2012. A network-based meta-population approach to model Rift Valley fever epidemics. *Journal of Theoretical Biology* 306 (2012), 129–144.

Forecasting the Flu: Designing Social Network Sensors for Epidemics

Huijuan Shao[1,2,*], K.S.M. Tozammel Hossain[5,*], Hao Wu[2,4,†], Maleq Khan[3], Anil Vullikanti[3], B. Aditya Prakash[1,2], Madhav Marathe[3], Naren Ramakrishnan[1,2]
[1]Department of Computer Science, Virginia Tech, Arlington, VA, USA
[2]Discovery Analytics Center, Virginia Tech, Arlington, VA, USA
[3]Biocomplexity Institute, Virginia Tech, Blacksburg, VA, USA
[4]Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA, USA
[5]Information Sciences Institute, University of Southern California, Marina Del Rey, CA, USA

ABSTRACT

Early detection and modeling of a contagious epidemic can provide important guidance about quelling the contagion, controlling its spread, or the effective design of countermeasures. A topic of recent interest has been the design of social network sensors, i.e., identifying a small set of people who can be monitored to provide insight into the emergence of an epidemic in a larger population. We formally pose the problem of designing social network sensors for flu epidemics and identify two different objectives that could be targeted in such sensor design problems. Using the graph theoretic notion of dominators we develop an efficient and effective heuristic for forecasting epidemics at lead time. Using six city-scale datasets generated by extensive microscopic epidemiological simulations involving millions of individuals, we illustrate the practical applicability of our methods and show significant benefits (up to twenty-two days more lead time) compared to other competitors.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Information systems** → *Data mining*; *Social networks*;

KEYWORDS

Social Network Sensors, Epidemic prediction, Health Informatics

ACM Reference Format:
Huijuan Shao[1,2,*], K.S.M. Tozammel Hossain[5,*], Hao Wu[2,4,†], Maleq Khan[3], and Anil Vullikanti[3], B. Aditya Prakash[1,2], Madhav Marathe[3], Naren Ramakrishnan[1,2]. 2018. Forecasting the Flu: Designing Social Network Sensors for Epidemics. In *Proceedings of ACM International Workshop on Epidemiology meets Data Mining and Konwledge Discovery (KDD epiDAMIK'18).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.475/123_4

[*]These authors contributed equally.
[†]Now working at Google Inc.

KDD epiDAMIK'18, August 2018, London, UK

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

Motivated by complicated public health concerns during the initial stages of a pandemic (other than just detecting if there is an epidemic at all) [11], public health officials are usually interested in the questions: Will there be a large disease outbreak? Or, has the epidemic reached its peak? These are important questions from a public health perspective [3]; the answers can help determine if costly interventions are needed (e.g., school closures), the strategies to organize vaccination campaigns and distributions, locations to prioritize efforts to minimize new infections, the time to issue advisories, and in general how to better engineer health care responses.

Given a graph and a contagion spreading on it, can we answer such questions by monitoring some nodes to get *ahead* of the overall epidemic? A social sensor is a set of individuals selected from the population which could indicate the outbreak of the disease under consideration, thus giving an early warning. Many existing methods for such detection problems typically give indicators which lag behind the epidemic. Recent work by Christakis and Fowler [5] has made some advances. They first proposed the notion of social network sensors for monitoring flu based on the friendship paradox: your friends have more friends than you do. They proposed a so-called 'Friend-of-Friend' approach to use the set of friends nominated by the individuals randomly sampled from the population as the social sensor. After implementing it among students at Harvard, Christakis and Fowler found that the peak of the daily incidence curve (the number of new infections per day) in the sensor set occurs 3.2 days earlier than that of a same-sized random set of students.

Figures 1 and 2 depict the results of experiments we did on two large contact networks—Oregon and Miami (see Table 1 for details)—using the SEIR model. We formed the sensor set using the approach given in [5] and measured the *average lead time* of the peaks for 100 runs (hence the results are robust to stochastic fluctuations). For the Oregon dataset, Fig. 1 shows that there is a lead time of 11 days on average for the peak in the sensor set with respect to the random set (see Fig. 1(c)). In contrast, for the Miami dataset, no lead time for the sensor set is observed (see Fig. 2(c)).

There may be several possible reasons for these inconsistencies. First, the 'Friend-of-Friend' approach implicitly assumes that the lead time always increases as we add more sensors into the set. Second, the lead time observation is assumed to be independent of the underlying network topology structures, which is clearly not

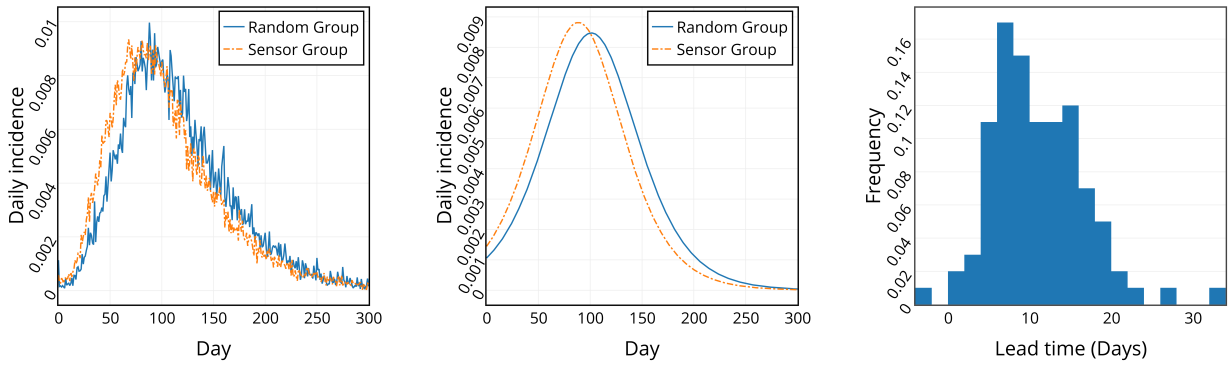


Figure 1: Illustration of the Friend-of-Friend approach [5] on the Oregon dataset. (a) True daily incidence curve (left), (b) fitted daily incidence curve with logistic function (middle), and (c) distribution of lead time over 100 experiments (right). Note that there is a non-zero lead time observed, i.e., the peak of the sensor curve occurs earlier than the peak of the curve for the random group.

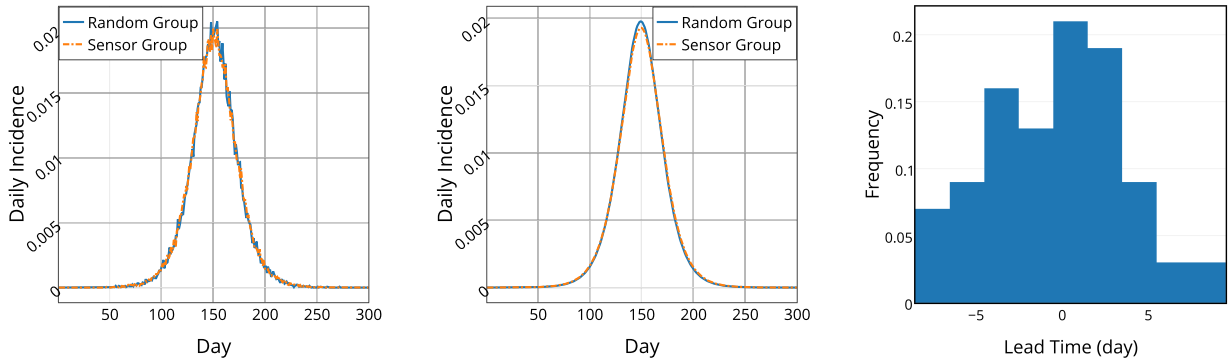


Figure 2: Illustration of the Friend-of-Friend approach on the Miami dataset. (a) True daily incidence curve (left), (b) fitted daily incidence curve with logistic function (middle), and (c) distribution of lead time over 100 experiments (right). Note that this experiment does not reveal any lead time.

the case. Finally, and most importantly, the work in [5] does not formally define the problem it is trying to solve, i.e., what objective does the sensor set optimize?

In this paper, we systematically formalize the problem of picking appropriate individuals to monitor and forecast the disease spreading over a social contact network. Our contributions are:

(1) We formally pose and study three variants of the sensor set selection problem.
(2) We provide an efficient heuristic based on the notion of graph dominators which solves one variant of the social sensor selection problem.
(3) We conduct extensive experiments on city-scale datasets based on detailed microscopic simulations, demonstrating improved lead time over competitors (including the Friend-of-Friend approach of [5]).
(4) We design surrogate/proxy social sensors using demographic information so that it is easy to deploy our approach in practice without knowledge of the full contact network.

2 EPIDEMIOLOGY FUNDAMENTALS

The most fundamental computational disease model is the so-called 'Susceptible-Infected' (SI) model where each individual (e.g. node in the disease propagation network) is considered to be in one of two states: Susceptible (healthy) or Infected. Any infected individual may infect each of its neighbors *independently* with probability β. Also, the SI model assumes every infected individual stays infected forever. For a clique of N nodes, the SI model can be characterized as:

$$\frac{dI}{dt} = \beta \times (N - I) \times I$$

where I is the number of infected nodes at time t. It is easy to prove that the solution for I is the logistic or sigmoid function, and its derivative (or the number of *new* infections per unit time) is symmetric around the peak.

The disease model that we use in this paper is the so-called SEIR model where a node in the disease propagation network is in one of *four* states: Susceptible, Exposed, Infected, and Recovered.The

dynamics of the SEIR model can be described as:

$$\frac{dS}{dt} = -\beta SI \quad \frac{dI}{dt} = \alpha E - \gamma I \quad \frac{dE}{dt} = \beta SI - \alpha E \quad \frac{dR}{dt} = \gamma I,$$

where S, E, I, and R denote the number of individuals in the corresponding states at time t, and $S + E + I + R = N$. Here β, α and γ represent the transition rates between the different states. Notice that since we are considering disease epidemics during a short period of time in this paper, we ignore the birth and death rates in the standard SEIR model here.

3 PROBLEM FORMULATION

Using the SEIR process, let $G = (V, E)$ be a social contact network where V and E represent the vertex set and edge set respectively. We use $f(S)$ to denote the probability that at least one vertex in the sensor set S gets infected, starting the disease spread from a random initial vertex. The most basic problem in such a setting is the *early detection* problem, in which the goal is to select the smallest sensor set S so that some vertices in S get infected within the first d days of the disease outbreak in the network G with probability at least ϵ (here, d and ϵ are given parameters)—this can be used to detect if there is an epidemic at all. This problem can be viewed as a special case of the detection problem in [10], and can be solved within a constant factor by a greedy submodular function maximization algorithm. As we show later, our optimization goal is *non-linear* and *not submodular*, and hence the approach in [10] can not be directly applied. Importantly, the early detection problem does not capture the more important issues about the disease characteristics of relevance to public health officials, and therefore we do not explore this further. For example, just detecting an infection in the population is generally not sufficient justification for doing an expensive intervention by public health officials (as the disease might not spread and may disappear soon). But knowing that the infection will still grow further and peak gives justification for robust infection control measures.

In our formulation, we use the term *epicurve* $I(t)$ to refer to the time series of number of infections by day. The *peak* of an epicurve is its maximum value, i.e., $\max_t I(t)$. Note that it is possible for an epicurve to have multiple peaks, but for most epidemic models in practice, the corresponding epicurves usually have a single peak. The derivative of the $I(t)$ with respect to t is called the *daily incidence* curve (number of new infections per day). The "time of peak" of the epicurve corresponding to the entire population is the time when the epicurve first reaches its peak, and is denoted by $t_{pk} = \text{argmax}_t\, I(t)$. Similarly, we use $t_{pk}(S)$ to denote the time-of-peak of the epicurve restricted only to a set S. The lead time of the epicurve peak for sensor set S compared to the entire population is then simply $t_{pk} - t_{pk}(S)$. The problem we study in this paper is:

(ϵ, k)-**Peak Lead Time Maximization (PLTM)**
Given: Parameters ϵ and k, network G, and the epidemic model
Find: A set of nodes S from G such that

$$S_{max} = \underset{S}{\text{argmax}}\, E[t_{pk} - t_{pk}(S)]$$
$$\text{s.t. } f(S) \geq \epsilon,\ |S| = k$$

Here, k is the budget, i.e. the required size of sensor set. Notice that we need the $f(S)$ constraint so that we only choose sets which have a minimum probability of capturing the epidemic—intuitively, there may be some nodes which only get infected infrequently, but the time they get infected during the disease propagation might be quite early. Such nodes are clearly not good 'sensors.'

4 PROPOSED APPROACH

Unfortunately, the peak of an epicurve is a high variance measure, making it challenging to address directly. Further, the expected lead time, $E[t_{pk} - t_{pk}(S)]$ is not non-decreasing (w.r.t. $|S|$) and non-submodular, in general. Hence we consider a different but related problem as an intermediate step. Let $t_{inf}(v)$ denote the expected infection time for node v, given that the epidemic starts at a random initial node. Then:

(ϵ, k)-**Minimum Average Infection Time (MAIT)**
Given: Parameters ϵ and k, network G, and the epidemic model
Find: A set S of nodes such that

$$S_{min} = \underset{S}{\text{argmin}} \sum_{v \in S} t_{inf}(v)/|S|$$
$$\text{s.t. } f(S) \geq \epsilon,\ |S| = k$$

Justification: In contrast to the peak, note that the *integral* of the epicurve restricted to S, normalized by $|S|$, corresponds to the *average infection time* of nodes in S, which is another useful metric for characterizing the epidemic. Further, if the epicurve has a sharp peak, which happens in most real networks and for most disease parameters, the average infection time is likely to be close to t_{pk}.

Approximating MAIT: The MAIT problem involves $f(S)$, which can be seen to be submodular, following the same arguments as in [7], and can be maximized using a greedy approach. However, the objective function — average infection time $\sum_{v \in S} t_{inf}(v)/|S|$ is non-linear as we keep adding nodes to S, which makes this problem challenging, and the standard greedy approaches for maximizing submodular functions and their extensions [8] do not work directly. In particular, we note that selecting a sensor set S which minimizes $\sum_{v \in S} t_{inf}(v)$ (with $f(S) \geq \epsilon$) might not be a good solution, since it might have a high average infection time $\sum_{v \in S} t_{inf}(v)/|S|$. We discuss below an approximation algorithm for this problem. For graph $G = (V, E)$, let $m = |E|$, $n = |V|$.

Lemma 1. *It is possible to obtain a bi-criteria approximation $S \subseteq V$ for any instance of the (ϵ, k)-MAIT problem on a graph $G = (V, E)$, given the $t_{inf}(\cdot)$ values for all nodes as input, such that $\sum_{v \in S} t_{inf}(v)$ is within a factor of two of the optimum, and $f(S) \geq c \cdot \epsilon$, for a constant c. The algorithm involves $O(n^2 \log n)$ evaluations of the function $f(\cdot)$.*

Proof. (Sketch) Let $t_{inf}(v)$ denote the expected infection time of $v \in V$, assuming the disease starts at a random initial node. Let B_{opt} be the average infection time value for the optimum; we can "guess" an estimate B' for this quantity within a factor of $1 + \delta$, by trying out powers of $(1 + \delta)^i$, for $i \leq \log n$, for any $\delta > 0$, since $B_{opt} \leq n$. We run $O(\log n)$ "phases" for each choice of B'.

Within each phase, we now consider the submodular function maximization problem to maximize $f(S)$, with two linear constraints: the first is $\sum t_{inf}(v)x(v) \leq B'k$ and $\sum_v x(v) \leq k$, where

$x(\cdot)$ denotes the characteristic vector of S. Using the result of Azar et al. [1], we get a set S such that $f(S) \geq c\mu(B')$, for a constant c, and $\sum_{v \in S} t_{inf}(v) \leq B'k$ and $|S| \leq k$, where $\mu(B')$ denotes the optimum solution corresponding to the choice of B' for this problem. If we have $|S| < k$, we add to it $k - |S|$ nodes with the minimum $t_{inf}(\cdot)$ values, which are not already in S, so that its size becomes k. Note that for the new set S, we have $\sum_{v \in S} t_{inf}(v) \leq 2B'k$, since the sum of the infection times of the nodes added to S is at most $B'k$.

Note that the resulting set S corresponds to one 'guess' of B'. We take the smallest value of B', which ensures $f(S) \geq c\epsilon$. It follows that for this solution S, we have $\sum_{v \in S} t_{inf}(v)/|S| \leq 2B_{opt}$ and $|S| = k$. The algorithm of Azar et al. [1] involves a greedy choice of a node each time; each such choice involves the evaluation of $f(S')$ for some set S', leading to $O(n^2)$ evaluations of the function $f(\cdot)$; since there are $O(\log n)$ phases, the lemma follows. □

Heuristics. Though Lemma 1 runs in polynomial time, it is quite impractical for the kinds of large graphs we study in this paper because of the need for a super-quadratic number of evaluations of $f(\cdot)$. Therefore, we consider faster heuristics for selecting sensor sets. The analysis of Lemma 1 suggests the following significantly faster greedy approach: pick nodes in non-decreasing $t_{inf}(\cdot)$ order till the resulting set S has $f(S) \geq \epsilon$. In general, this approach might not give good approximation guarantees. However, when the network has "hubs", it seems quite likely that the greedy approach will work well. However, even this approach requires repeated evaluation of $f(S)$, and can be quite slow. The class of social networks we study has the following property: nodes v which have low $t_{inf}(v)$ are usually hubs and have relatively high probability of becoming infected. This motivates the following simpler and much faster heuristic, referred to as the **Transmission tree (TT) based sensors** heuristic:

(1) generate a set $\mathcal{T} = \{T_1, \ldots, T_N\}$ of dendrograms; a dendrogram $T_i = (V_i, E_i)$ is a subgraph of $G = (V, E)$, where V_i is the set of infected nodes and an edge $(u, v) \in E$ is in E_i iff the disease is transmitted via (u, v);
(2) for each node v, compute d_v^i, which is its depth in T_i, for all i, if v gets infected in T_i;
(3) compute $t_{inf}(v)$ as the average of the d_v^i, over all the dendograms T_i, in which it gets infected;
(4) discard nodes v with $t_{inf}(v) < \epsilon_0$, where ϵ_0 is a parameter for the algorithm;
(5) order the remaining nodes $v_1, \ldots, v_{n'}$ in non-decreasing $t_{inf}(\cdot)$ order (i.e., $t_{inf}(v_1) \leq t_{inf}(v_2) \leq \ldots \leq t_{inf}(v_{n'})$)
(6) Let $S = \{v_1, \ldots, v_k\}$

We also use a faster approach based on dominator trees, which is motivated by the same greedy idea. We referred to it as the **Dominator tree (DT) based sensors** heuristic:

(1) generate dominator trees corresponding to each dendrogram;
(2) compute the average depth of each node v in the dominator trees (as in the transmission tree heuristic);
(3) discard nodes whose average depth is smaller than ϵ_0;
(4) order nodes based on their average depth in the dominator tree, and pick S to be the set of the first k nodes.

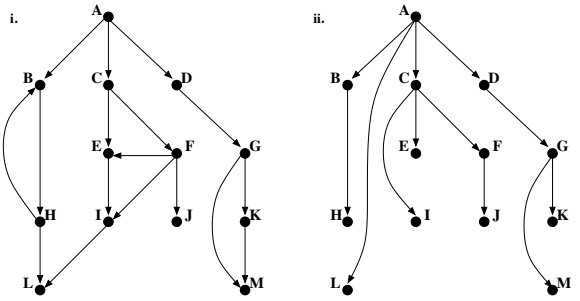


Figure 3: (i) An example graph and (ii) its dominator tree. In practice, the dominator will have a significantly reduced number of edges than the original graph.

Formally, the dominator relationship is defined as follows. A node x dominates a node y in a directed graph iff all paths from a designated start node to node y must pass through node x. In our case, the start node indicates the source of the infection or disease. Consider Fig. 3 (left), a schematic of a social contact network. All paths from node A (the designated start node) to node H must pass through node B, therefore B dominates H. Note that a person can be dominated by many other people. For instance, both C and F dominate J, and C dominates F. A node x is said to be the unique immediate dominator of y iff x dominates y and there does not exist a node z such that x dominates z and z dominates y. Note that a node can have at most one immediate dominator, but may be the immediate dominator of any number of nodes. The dominator tree $D = (V^D, E^D)$ is a tree induced from the original directed graph $G = (V^G, E^G)$, where $V^D = V^G$, but an edge $(u \rightarrow v) \in E^D$ iff u is the immediate dominator of v in G. Fig. 3 (right) shows an example dominator tree.

The computation of dominators is a well studied topic and we adopt the Lengauer-Tarjan algorithm [9] from the Boost graph library implementation. This algorithm runs in $O((|V|+|E|)\log(|V|+|E|))$ time, where $|V|$ is the number of vertices and $|E|$ is the number of edges.

5 EXPERIMENTAL RESULTS

Our experimental investigations focus on addressing the following questions:

(1) How do the proposed approaches perform when forecasting the epidemic in terms of the lead time?
(2) How large should our sensor set size be?
(3) How many days are necessary to observe a stable lead time?
(4) What is the predictive power of the sensor set in estimating the epidemic curve over the full population?
(5) Is it possible to employ surrogates for sensors?

Table 1 shows some basic network statistics of the datasets we used in our experiments. The Oregon AS (Autonomous System) router graph is an AS-level connectivity network inferred from Oregon route-views [4]. Although this dataset does not relate to epidemiological modeling, we use it primarily as a testbed to understand how (and if) graph topology affects our results due to the relatively small size and neat graph structure. The rest of the datasets are synthetic but realistic social contact networks (see [2, 6]) for six large cities

Table 1: Characteristics of datasets used in the experiments.

Dataset	Nodes	Avg. deg	Max deg
Oregon	10,670	4.12	2,312
Miami	2,092,147	50.38	425
Boston	4,149,279	108.32	437
Dallas	5,098,598	113.10	477
Chicago	9,047,574	118.83	507
Los Angeles	16,244,426	113.08	463
New York	20,618,488	93.14	464

in the United States. These six US city datasets are generated with specific aim at modeling epidemics in human populations.

In our experimental study, we evaluated our two proposed approaches: the transmission tree based heuristic and the dominator tree based heuristic. For comparison, we also implemented two strategies as baseline methods: (i) a **Top-K high degree sensors** heuristic used in [5] where a set $P \subseteq V$ is first sampled and for each $v \in P$ its K neighbors with largest degree are selected, and (ii) a **Weighted degree (WD) sensors** heuristic, which is similar to the previous heuristic except that the K neighbors are chosen based on largest weighted degree. The weight we use here is the durations of the activities indicated by edges of the graphs in the datasets mentioned in Table 1. However, since we don't have these weights for the Oregon dataset, we will omit the results of the WD sensor heuristic on the Oregon dataset.

Our primary figure of merit is the lead time, calculated as follows. For each run of the disease model in a social contact network, we fit a logistic function curve to the cumulative incidence of the chosen sensor set and a random sampled set from V. Here, we use the random sampled set to represent the entire population for the large city-level datasets we used in our experiments. (It is usually impossible to track the entire population in practice.) We then derive daily incidence curves for both the sensor set and the random set (we will refer to this set as random set in the rest of this paper). Let t_s and t_r represent the peak times of the daily incidence curves for the sensor and random sets respectively, and the lead time is defined as $\Delta t = t_r - t_s$.

For all the experiments in this section, the parameters for the epidemic simulations are set as follows unless specified. We set $\epsilon = 0.8$ (see the definitions of the PLTM and MAIT problems) and flu transmission rate to be 4.2×10^{-5} for the SEIR disease model. The size for the sensor set and random set (k) is 5% of the entire population, and the epidemic simulations start with five randomly infected vertices in the networks. All the results were obtained by averaging across $1,000$ independent runs.

5.1 Performance of predicted epidemic lead time

In this experimental study, we set the flu transmission rate to 0.05 for the SEIR model in the Oregon dataset due to its relatively small size compared to the Miami dataset. Fig. 4 depicts the daily incidence curves of the four sensor selection heuristics and the random set on Oregon and Miami datasets, and Fig. 5 describes the corresponding peak time of the daily incidence curves shown in Fig. 4. As we can see from these figures, on the Oregon dataset, the performance of the proposed heuristics and baseline heuristics is comparable where they both predict the peak of the epicurves about five days earlier when compared to the ground truth. However, on the Miami dataset, the proposed TT and DT heuristic approaches give a much larger lead time, around 10 days, compared to the about two-day and almost zero day lead time in the WD and Top-K baseline heuristics. This is because, as described earlier, our approaches are precisely designed to try to pick vertices with early expected infection time from the disease propagation network as social sensors. We also study whether the number of the initial infected vertices will affect the predicted lead time. Table 2 shows the predicted lead time of the two proposed and the two baseline heuristics for 1, 5, and 10 initial infected vertices in the epidemic simulations. As the results in this table show, the number of initial infected vertices would not have too much impact on the predicted lead time.

5.2 How many sensors to choose?

Since we have already demonstrated the influences of the network topology on social sensor selection strategies, we will put the Oregon dataset aside, and focus on the social contact network datasets for US cities in the rest of the experiments. An interesting conundrum is the number of sensors to select in a design. Fig. 6 depicts the mean lead time and the inverse of variance-to-mean ratio of the lead time v.s. the sensor size for the Miami datasets. The results show that the variance of the lead time estimate is high for small size of sensor sets and decreases as the sensor set size increases. This suggests a natural strategy of scaling the lead time against the variance, thus helps establish a sweet spot in the trade-off. This variance-to-mean ratio is also known as the *Fano factor*, which is widely used as an index of dispersion. In the result for the Miami dataset, there is a clear peak in the figure of the inverse of variance-to-mean ratio, which suggests a suitable size of sensors to pick.

5.3 Empirical study on stability of lead time

In this experiment, we study the stability of the estimated lead time as we observe more data on the sensor group when the number of monitoring days increases. As is well known, the cumulative incidence curve of flu epidemics can be modeled by a logistic function where the dependent and independent variables are the flu cumulative incidence and the time of the epidemic (days in our context). Here, we vary our flu epidemic simulation time from 2 days to 300 days on the Miami dataset, estimate cumulative incidence curves (with logistic function) for both the sensor and the random set based on the simulated cumulative flu incidence data, and then compute the lead time. Fig. 7 shows the lead time vs. the flu epidemic simulation time. As we can see from this figure, the estimated lead time fluctuates a lot when the simulation time is short and stabilizes at around 12 days when the epidemic simulation time is more than around 80 days. Such results provide some insights for public health officials on how much epidemic data they should collect in order to make an accurate estimation of the flu outbreak from the time domain perspective.

5.4 Predicting population epidemic curve from sensor group epidemic curve

In this experiment, we study the relationship between the flu cumulative incidence curve of sensor and that of random group. As

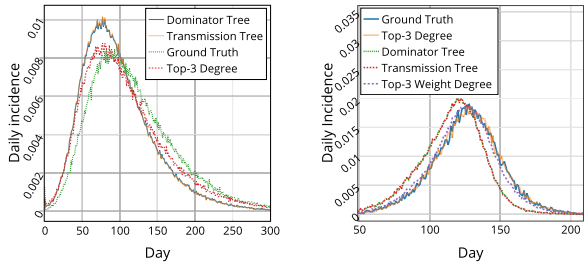


Figure 4: Daily incidence of sensor sets selected by the heuristic approaches compared to the true daily incidence in the simulated epidemic on (a) Oregon dataset (left), (b) Miami dataset (right).

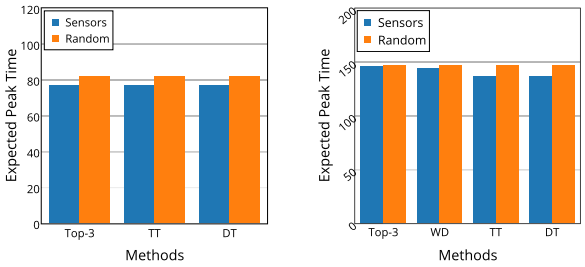


Figure 5: The expected peak time of the daily incidence curve on (a) Oregon dataset (left), (b) Miami dataset (right). Here Top-3, WD, TT, and DT denote Top-3 high degree, Top-3 weighted degree, Transmission tree based, and Dominator tree based heuristic respectively.

Table 2: Comparison of the lead time across four different social sensor selection heuristics when the number of initial infected vertices vary.

Dataset	Seed	Lead time			
		Top-K degree	Weight degree	Transmission tree	Dominator tree
Oregon	1	13.13	n/a	10.10	9.91
	5	8.85	n/a	7.93	7.75
	10	11.00	n/a	8.63	8.55
Miami	1	0.29	3.38	10.46	10.08
	5	0.39	3.41	10.15	10.19
	10	0.62	3.41	10.13	10.13

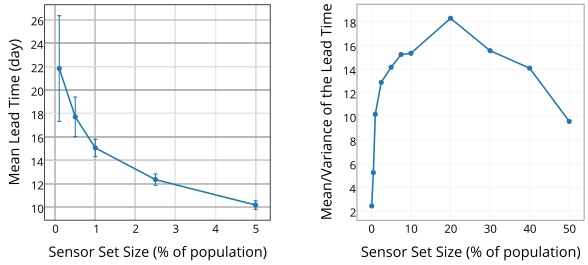


Figure 6: Mean lead time (left) and inverse of variance-to-mean ratio (right) v.s. the sensor size for the Miami dataset. When sensor set size is less than 1.0% of the entire population we observe higher (good) lead time, but also with high variances. Scaling the mean lead time by the variance, i.e., the reciprocal of the Fano factor, shows a clear peak with the sensor set size at approximately 20% of the population, the position where we can obtain substantial gains in lead time with correspondingly low variances.

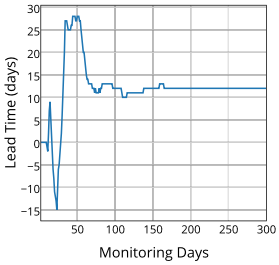


Figure 7: Stability of the lead time estimation. The estimated the lead time fluctuates initially. As the number of monitoring days increases, it stabilizes quickly.

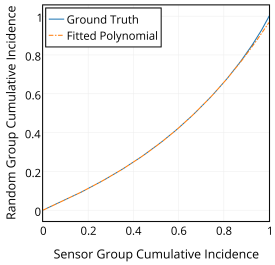


Figure 8: Predicting cumulative incidence of random group with sensor group for the Miami dataset.

we mentioned before, we use random set to represent the entire population since it is usually quite difficult to characterize the entire population in practice when the dataset is quite large. We try to estimate a polynomial regression model with degree of three where the observed cumulative incidence of the sensor group serves as predictor and that of the random group serves as responses. Here, the sensor group is selected by the dominator tree heuristic from the Miami dataset. Over the 300 simulated days, we use the data of the first 150 days to estimate our polynomial regression model, and make predictions of the cumulative incidence of random group for the rest of the 150 days. Fig. 8 shows the fitted polynomial regression model compared to the true relation curve of the flu cumulative incidences between sensor group and random group. As we can see from this figure, the polynomial regression model with degree of three could capture the relationship between the cumulative incidences of random group and sensor group quite well, which can help us predict the epidemic curve of entire population with epidemic data collected from the sensor group.

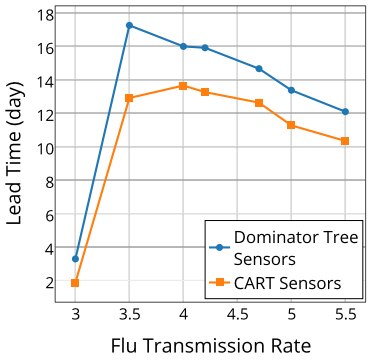


Figure 9: Mean lead times estimated with surrogate sensor set S'' and dominator tree based social sensors for various flu transmission rates.

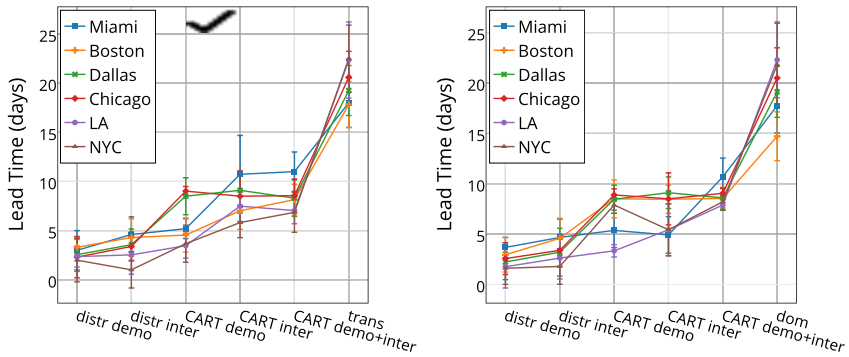


Figure 10: The lead time of transmission tree based (left) and dominator tree based (right) sensor selection strategies using different combinations of individual demographic and interaction information on Miami, Boston, Dallas, Chicago, Los Angeles and New York City datasets.

5.5 Surrogates for social sensors

In reality, the structures of large scale social contact networks are usually unknown or difficult to obtain, which makes it difficult to directly apply our proposed methods as we have done thus far. In order to make the proposed approaches deployable and solve realistic public health problems, we now relax this key assumption, and develop a *surrogate* approach to select social sensors. In this case, the policy makers can implement their strategies without detailed (and intrusive) knowledge of people and their activities. Surrogates are thus an approach to implement privacy-preserving social network sensors.

The key idea of our surrogate approach is to utilize the demographic information. Here, we use the Miami dataset as an example to explain our surrogate approach. We extracted the following 16 demographic features from the Miami dataset: age, gender, and income; number of meetings with neighbor nodes; total meeting duration with neighbor nodes; number of meetings whose durations are longer than 20000 seconds; number of meetings of types 1–5; and percent of meetings of types 1–5. The meeting types of 1–5 refer to home, work, shop, visit, and school, respectively. To select surrogate sensors using demographic information, we use classification and regression trees (CART); any other supervised classification algorithm can also be substituted here. The 16 attributes mentioned above are used as independent variables in our CART model, and the response variable is binary to indicate whether a person should be selected as a sensor or not. In order to learn the CART model, we create the training data as follows. We choose 0.1% of the entire population ($\approx$ 2000) from the US city dataset with our proposed heuristics as the training data with positive responses (social sensors), and choose another 0.1% randomly as the training data with negative responses (not social sensors). Then, separate CART models were learned to select the surrogate sensor set S' for each transmission rate ranging from 3.0×10^{-5} to 5.5×10^{-5} with a step size of 5×10^{-6}. Such transmission rates are the typical values used in various flu epidemic studies. Among all the surrogate sensors chosen by each of these CART models, we choose the common individuals across all the CART models as the final surrogate sensor set S''.

Fig. 9 compares the estimated lead time between the surrogate sensor set S'' and the sensor set selected by the dominator tree heuristic for various flu transmission rates. As we can see from this figure, although the surrogate sensor set S'' does not perform as well as the proposed dominator tree based sensor set, it still provides a significant lead time, which is good enough to give early warning to public health officials for the potential incoming flu outbreak. Most important, since the CART based surrogate sensor approach does not require the information of the social contact network structures, it is easy to implement and deploy in reality compared to the transmission tree and dominator tree based heuristic approaches. This makes it a promising candidate for predicting flu outbreaks for public health officials.

5.6 What information should be used to select surrogate sensors?

Notice that in the last section, when we select the surrogate sensors, both demographic (e.g. age of individuals) and interaction (e.g. total meeting duration and meeting types with neighboring individuals) information is taken into account. However, which kind of information is more important in terms of estimating the lead time of flu epidemics? In this experiment, we focus on all our social contact network datasets for large US cities, i.e., Miami, Boston, Dallas, Chicago, Los Angeles, and New York. For each city, we selected the surrogate sensor set and the random set with the fixed size of 10, 000. The sensor set was selected with the following six strategies: 1) using empirical distributions of demographic information (distr demo); 2) using empirical distributions of interaction information (distr inter); 3) using CART with demographic information (CART demo); 4) using CART with interaction information (CART inter); 5) using CART with both demographic and interaction information (CART demo+inter); 6) using transmission tree or dominator tree based heuristic (trans or dom). We computed the lead time for each of the six surrogate sensor selection strategies mentioned above, and the results were averaged across 100 independent runs. Fig. 10 shows the lead time of the different approaches over the six US city datasets. As we can see from the figure, our proposed approaches (CART based approaches and transmission/dominator

tree based approaches) outperform the two baseline methods (distr demo/inter), and in general, as more information is taken into account, the larger estimated lead time could be achieved (since the transmission/dominator tree based heuristics assume known social contact network structures, they could be thought of possessing the most information about epidemics). Furthermore, the individual interaction information seems to be more important than the demographic information from the perspective of obtaining larger lead time.

6 DISCUSSION

The most closely related work to ours is Christakis and Fowler [5], where a simple heuristic that monitors the friends of randomly chosen individuals from a social network as sensors was adopted to achieve early detection of epidemics. However, they only demonstrated their proposed approach on a relatively small social network, e.g. a student network from Harvard College. As we have shown earlier, their friend heuristic fails on large social contact networks of US cities. We have also demonstrated that although the Christakis and Fowler's approach works well over small networks like the Oregon dataset, it provides almost no lead time over large scale social contact networks like the Miami dataset. To explain why the proposed social sensor selection heuristics work better, we start from analyzing the structures of the disease propagation networks. Comparing the graph statistics of the Oregon dataset with the Miami dataset shown in Table 1, we can observe that the graph in the Oregon dataset has a quite different topology structure from the graphs in the Miami datasets. The graph in the Oregon dataset has relatively small average degree but very large maximum degree, which indicates this graph has a star-like topology where few of the central vertices have very large degrees. On the other hand, many vertices in the graphs of the Miami datasets have large degrees, and they spread all over the entire graph. Thus, for the top-K degree based sensor selection approach, it is relatively easy to include the central vertices with high degrees into the sensor set in the Oregon dataset, but for the transmission tree and dominator tree based approaches, whether the high degree vertices are included into the sensor set will heavily depend on the choices of initial seeds of the epidemics in the Oregon network. Such central vertices with high degree are usually very important for the epidemics in such star-like networks, which explains why the top-K degree approach works better than the transmission tree and dominator tree approaches. On the contrary, in the Miami dataset, the total number of vertices is large, and it is quite difficult for the top-K degree approach to select sensors that could represent the entire graph only based on local friend-friend information. However, the transmission tree and dominator tree based sensor selection strategies take the global epidemic spread information into account, which chooses the sensor set that could represent the entire graph. That's why they perform better in terms of the lead time than the top-K degree based approach on the large simulated US city networks. The interesting insight revealed by such results is that the network topology must be considered when designing social sensor selection strategies. The results also demonstrate that the proposed TT and DT based sensor selection heuristics are more robust to the underlying network topologies, and thus more suitable to be deployed in practice, such as monitoring and forecasting epidemics in large cities.

7 CONCLUSION

In this paper, we studied the problem of predicting flu outbreaks with social network sensors. Compared to previous works, we are the first to systematically formalize and study this problem. By leveraging the graph theoretic notion of dominators, we developed an efficient heuristic to select good social sensors to forecast the flu epidemics when the structure of flu propagation network is known. Redescription of the dominator property in terms of demographic information enables us to develop a truly implementable and deployable strategy to select surrogate social sensors to monitor and forecast flu epidemics, which will benefit public health officials and government policy makers.

ACKNOWLEDGMENTS

This work is supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, by the National Science Foundation via grants DGE-1545362, IIS-1633363, and by the Army Research Laboratory under grant W911NF-17-1-0021. The US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, NSF, Army Research Laboratory, or the U.S. Government.

This paper is also based on work partially supported by the NEH (HG-229283-15), NSF CAREER (IIS-1750407), ORNL (Task Order 4000143330), and a Facebook faculty gift

REFERENCES

[1] Y. Azar and I. Gamzu. 2012. Efficient Submodular Function Maximization under Linear Packing Constraints. In *ICALP*.
[2] C. Barrett, D. Beckman, M. Khan, V.S. Anil Kumar, M. Marathe, P. Stretz, T. Dutta, and B. Lewis. 2009. Generation and analysis of large synthetic social contact networks. In *Winter Simulation Conference*.
[3] CDC. 2012. Flu Activity During the 2012-2013 Season. http://www.cdc.gov/flu/about/season/flu-season-2012-2013.htm
[4] Q. Chen, H. Chang, R. Govindan, and S. Jamin. 2002. The Origin of Power Laws in Internet Topologies Revisited. In *INFOCOM '02*. IEEE, 608–617.
[5] N.A. Christakis and J.H. Fowler. 2010. Social Network Sensors for Early Detection of Contagious Outbreaks. *PloS one* 5, 9 (2010), e12948.
[6] S. Eubank, H. Guclu, V. S. Anil Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature* 429, 180-184 (2004).
[7] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Conference of the ACM Special Interest Group on Knowledge Discovery and Data Mining*. ACM Press, New York, NY.
[8] Andreas Krause and Carlos Guestrin. 2008. Beyond Convexity - Submodularity in Machine Learning. In *ICML*. http://submodularity.org/
[9] T. Lengauer and R. Tarjan. 1979. A fast algorithm for finding dominators in a flowgraph. *ACM Trans. Program. Lang. Syst.* 1, 1 (1979), 121–141.
[10] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie S. Glance. 2007. Cost-effective outbreak detection in networks. In *KDD*. 420–429.
[11] P. Nsubuga, M. White, and S. Thacker. 2006. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions. In *Disease Control Priorities in Developing Countries, World Bank*. http://www.ncbi.nlm.nih.gov/books/NBK11770/.

A TOPOLOGICAL DATA ANALYSIS APPROACH TO INFLUENZA-LIKE-ILLNESS

Joao Pita Costa[1 2 3], Primož Škraba[3 4], Daniela Paolotti[5] and Ricardo Mexia[6]

***Abstract*—Influenzanet is an online automated system to monitor the activity of influenza-like-illnesses (ILI) with the aid volunteers through the internet. The discussion in this paper has focus on the topological analysis of the Influenzanet dataset, examining the structure of that data to provide insights on the behaviour of the ILI season and comparing ILI seasons. The general approach performs a qualitative analysis based on the topology of the curves of the time series generated by each ILI season. It provides a way to test agreement at a global scale arising from local models. We also show the complementary potential of this qualitative method to quantitative methods such as Fourier analysis and dynamical time warping.**

***Keywords*—digital epidemiology, influenza-like-illness, influenzanet, persistent homology, computational topology, persistence diagram.**

I. INTRODUCTION

Due to the pandemic potential of influenza, a complete knowledge of the development of each ILI season is a public health priority. In this paper we contribute to that aim by discussing the problem of comparing influenza seasons throughout the years (selected countries: Portugal and Italy) based on the topological behavior of the curve of the time series of incidence in the population. We also discuss the identification of recurrence that would correspond to patterns in the influenza season. To do that we recur to the novel methods of topological data analysis [TDA], providing us with the persistent topological features that describe the structure of the data. The basic technique encodes topological features of a given point cloud by diagrams representing the lifetime of those topological features (see Figure 1). A good introduction to topological data analysis can be found in [2]. Topological methods on data have seen application to the study of the influenza viral evolution in [3] and other public health priorities such as diabetes [5] or cancer [7]. Our goal in this paper is to analyze the Influenzanet data using persistent homology (i.e. *persistence*), identifying persistent topological features relevant to the digital epidemiology study. The *Influenzanet* system monitors the activity of *influenza-like-illness* [ILI] in Europe with the aid of online volunteers. It has been operational in Portugal since 2005, and in Italy since 2008. Influenzanet obtains its data directly from the population, contrasting with the traditional system of sentinel networks of mainly primary care physicians [8]. Influenzanet was shown to be a fast and flexible monitoring system whose uniformity allows for direct comparison of ILI rates between countries [13]. In this paper we look at the overall structure of several influenza seasons as well as their evolution in Portugal and Italy. In particular, this provides a way to test agreement at a global scale arising from standard local models. The method for comparing time series data through TDA, is innovative in the context of the analysis of ILI seasons. It differs from other approaches by providing us with a tool that is independent of the different sizes of the samples collected in each country, comparing the shape of the data generated in each ILI season between countries. We will compare it with dynamic time warping [DTW], that can also compare the time series based on their behaviour, independent from the variations in time. A complementary study is to look for periodicity in the ILI season. The usage of TDA for the analysis of time series was explored in [9] towards the quantification of periodicity and identification of periodic signals in gene expression. The method infers high-dimensional structure from low-dimensional representations and studies properties of a continuous space by the analysis of a discrete sample of it, assembling discrete points into global structure. Similarly, using TDA to analyze the input time series data, after a delay embedding of the time series in R^2 as in [9], we can study periodicity in the Influenzanet data [10], or compare the persistent features of the curves generated by that data [12]. We compare the results with those of Fourier analysis, the standard quantitative analysis of periodicity in the data.

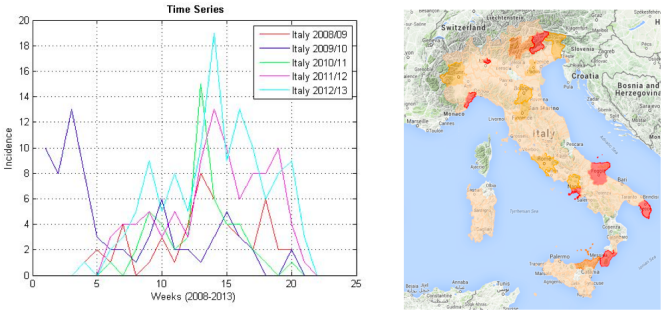


Fig. 1. The curves of the time series of ILI incidence in Italy during the ILI seasons of 2008/09-2012/13 (on the left); a screenshot of the influenzanet system in Italy, in May 2015 (on the right), where the infection level based on reported symptoms goes from ocher (min) to red (max).

II. TOPOLOGICAL ANALYSIS OF EPIDEMIOLOGICAL DATA

Given the time series of ILI incidence defined by pairs (country, year), we aim to compare them through the analysis of the persistence of topological features. In particular, we

[1]University of Rijeka, Croatia; [2]Quintelligence, Slovenia; [3]Institute Jožef Stefan, Slovenia; [4]University of Primorska, Slovenia; [5]ISI Foundation, Italy; and [6]Instituto Nacional de Saúde Dr. Ricardo Jorge, Portugal

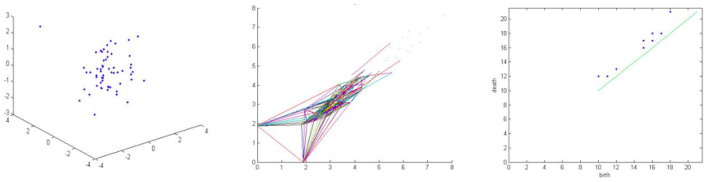


Fig. 2. The pipeline for the computation of topological data analysis for the time series of Italy for 2009/10: the given pointcloud of the input data (on the left); the Viatoris-Rips complex approximating the space of the pointcloud (in the center); and the correspondent persistence diagram encoding the lifetime of the topological features (on the right).

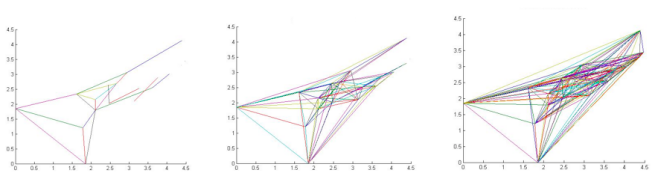


Fig. 3. The filtration of the simplicial complex at several levels varying according a parameter r for the input time series of Italy in the ILI season of 2009/2010: $r = 2$ (on the left); $r = 3$ (in the center); $r = 5$ (on the right).

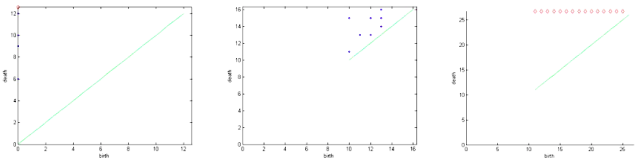


Fig. 4. The persistence diagrams for the input time series of Italy in the ILI season of 2009/10: dimension 0 (on the left); dimension 1 (in the center); dimension 2 (on the right). The red circles mean that the lifetime of the considered features does not end.

embed the data in higher dimensions, compute persistence, and distinguish data noise and outliers. In that, we get a perspective of that space under different scales, where small features will eventually disappear. The approach in [14] using the Takens Delay Embedding Theorem permits us to project the time series data onto a n-dimensonal space and from it construct a persistence diagram corresponding to the time series of influenza incidence. The Mahalanobis distance is a measure of the distance between a point P and a distribution D, widely used in cluster analysis and classification techniques. As persistence is independent from the metric used, in this paper we consider the Mahalanobis metrics on the space to construct a *simplicial complex* (i.e. a combinatorial approximation of the space based on points, line segments, triangles, and their n-dimensional counterparts [2]) within the TDA analysis pipeline, represented in Figure 2. The complexity of computations grows fast with the rise of input data due to the usage of these topological structures. For the sake of efficiency we have used several methods to preprocess the given data. The computation of persistence is fed by the time series data embedded in higher dimensions and provided as a distance matrix to the algorithm that computes the persistent features encoded into (persistence) diagrams. The images in Figure 2 show the cloud of input data points, the corresponding simplicial complex (a Vietoris-Rips complex), and the corresponding persistence diagram for dimension 1 (where the information on topological cycles is captured). The computation of the persistence diagrams is done using Ripser [1], an open source persistent homology software that can output a text file with a list of birth and death times corresponding to the measure of persistence, fully describing the persistence diagram. We also used an alternative open source tool, Perseus [6], whenever we needed to control parameters of persistence computations (eg. step size, number of steps or initial threshold distance). The input structure is given as a symmetric distance matrix where the entries come from pairwise distances between points in a given point cloud. In the Figure 3 we can see the 3-step construction of the Vietoris-Rips complex that will provide us with the persistence diagram encoding the topological information of the Influenzanet data. These topological tools complement the information obtained by classical data analysis.

III. COMPARING ILI SEASONS

When comparing two time series that may vary in time or speed it is reasonable to apply DTW to measure the similarity between the temporal sequences. Doing so, we are able to align the time series to enable comparison between seasons. In this study we compared each pair (country, year), obtaining the respective measures in the table of Figure 5, with highlighted largest and smallest values. We compare these values with those also in Figure 5 coming from comparing persistence diagrams each of which corresponding to an influenza season in either Portugal or Italy (for the embedding using and comparing Euclidean and Mahalanobis metrics). Bottleneck distance is a standard technique of TDA that permit us to measure the pairwise distance between persistence diagrams at each dimension. The distance value between two persistence diagrams in the tables of Figure 5 was calculated using the persistence landscapes toolbox [4] to compute the distance between diagrams. Persistence Landscapes generalizes

SW Persistence
Mahalanobis

		Italy				
		2008	2009	2010	2011	2012
Portugal	2008	0.69054	0.67536	0.51681	0.66377	0.58568
	2009	0.53593	0.52165	0.37944	0.35339	0.38572
	2010	0	**0.031433**	0.24607	0.27187	0.1758
	2011	0.1758	0.18699	0.28006	0.32567	0.27187
	2012	**1**	0.98653	0.81235	0.90479	0.86585

DTW

		Italy				
		2008	2009	2010	2011	2012
Portugal	2008	0.87255	**1**	0.69608	0.57843	0.40196
	2009	0.78431	0.31373	0.73529	0.72549	0.91176
	2010	0.19608	0.47059	0.13725	0.14706	0.42157
	2011	**0.17647**	0.47059	0.2451	0.27451	0.5098
	2012	0.22549	0.54902	0.26471	0.38235	0.53922

Fig. 5. Comparing the ILI seasons of Portugal and Italy during 2008/09 up to 2012/13: the distance tables for the TDA with Euclidean metrics (on the top), the TDA with Mahalanobis metrics (on the center), and the dynamic time warping (on the bottom).

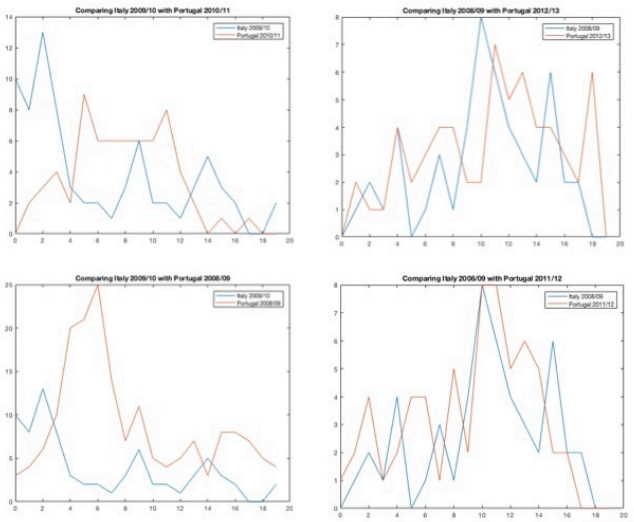


Fig. 6. Comparing the ILI seasons of Portugal and Italy during 2008/09-2012/13: selected plots of time series to compare the results in the topological data analysis and the dynamical time warping.

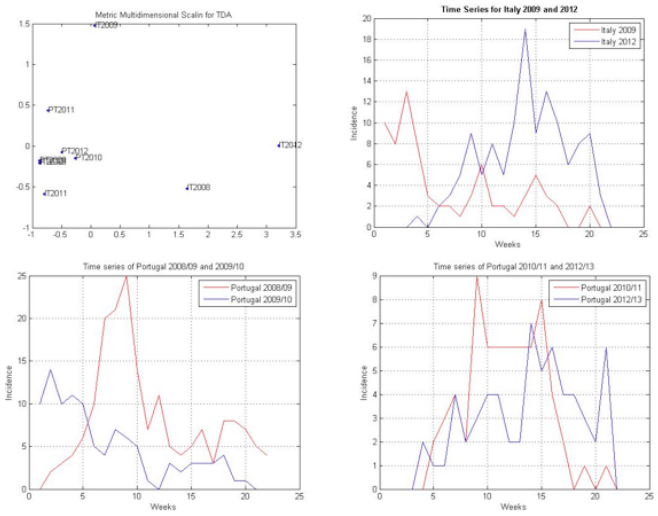


Fig. 7. Comparing the ILI seasons of Italy and Portugal during 2008-2013 using metric multidimensional scaling (on the upper left) to identify: the outlier ILI seasons of Italy 2009/10 and 2012/13, with time series plotted for analysis and interpretation (on the upper right); the close ILI seasons of Portugal 2008/09 and 2009/10 (on the lower left); and the ILI seasons of Portugal 2010/11 and 2012/13, close to the diagonal (on the lower right).

bottleneck distance and will be used in further research to get deeper insights from these comparisons. The tables in Figure 5 represent the comparison between the TDA and DTW analyses of ILI incidence in Italy and Portugal for the ILI seasons of 2008/09 up to 2012/13. This data was normalized by maximum distance (i.e., using normalized = $(x - min(x))/(max(x) - min(x))$), enabling the comparison of $[0, 1]$ values in such non homogeneous data. When comparing the distances obtained by DTW and TDA we can see that these two methods look at different features of the data and thus the different results obtained. The plots in Figure 6 represent time series for the selected ILI seasons. They allow us to compare the different data analyses methods used in this study. When comparing the distances between the ILI seasons in Italy and Portugal, the TDA often disagrees with DTW (see Figure 6). The closest ILI seasons according to TDA are those of Italy 2009/10 and Portugal 2010/11, where the TDA has a lowest value of 0.0314 (due to the higher similarity of peaks) while the DTW has an average value of 0.4706. The closest ILI seasons according to DTW are those of Italy 2008/09 and Portugal 2011/12, where the DTW has a lowest value of 0.1765 (describing the similar behavior of the curves) while the TDA analysis has also a low value of 0.1758. We used multidimensional scaling as in Figure 7 to identify outliers for each of the three methods within the ILI seasons analyzed in this study. TDA provides a qualitative analysis of the time series of ILI incidence, looking in particular at the peaks and dramatic changes. In that perspective, the time series of Italy 2009/10 and 2012/13 plotted in Figure 7 describe very different ILI seasons with very different peaks. On the other hand, the ILI seasons of Portugal 2008/09 and 2009/10 are identified being very close with very similar peaks, although the behavior of the curve being different (it is worth mentioning that these were seasons influenced by the pandemic H1N1/09 virus). The knowledge on secondary attack rates in the influenza season is of importance to access the severity of the seasonal epidemics of the virus, estimated recently with information extracted from social media in [15]. Here lies a strong point of TDA where it can provide relevant contribution complementing other methods. The persistence diagrams of Figure 8, correspondent to the identified ILI seasons of Italy 2009/10 and 2012/13, and Portugal 2008/09 and 2009/10. They encode the lifetimes of the topological features of the curves of the time series of those seasons. Persistence diagrams are a clear and practical tool that allows us the detection of outliers and to capture the qualitative features of the dynamics of the system. These ideas provide a new approach to the analysis of the seasons in the epidemiology of Influenza.

IV. LOOKING FOR PERIODICITY IN THE INFLUENZANET DATA

Fourier analysis is widely used to identify patterns in a time series. In this section we discuss how the qualitative data analysis of TDA can complement the quantitative information provided by the Fourier analysis. In Figure 3 we can see the plot of the two time series and their correspondent Fourier transform. We used the time series of ILI incidence in Portugal and Italy for the ILI seasons of 2008-2013, representing non-homogeneous data. We computed the Fourier transform for each pair of time series (country, year) to compare the ILI seasons of Portugal and Italy, as in [11], confronting the quantitative methods of the Fourier analysis with the qualitative methods of TDA. TDA can also be used to look for periodicity in Influenzanet data, following the work in [14], to identify recurrent behaviours within selected influenza seasons. Barcodes and the correspondent persistence diagrams, seen as multi-scale signatures encode the lifetime of topological features

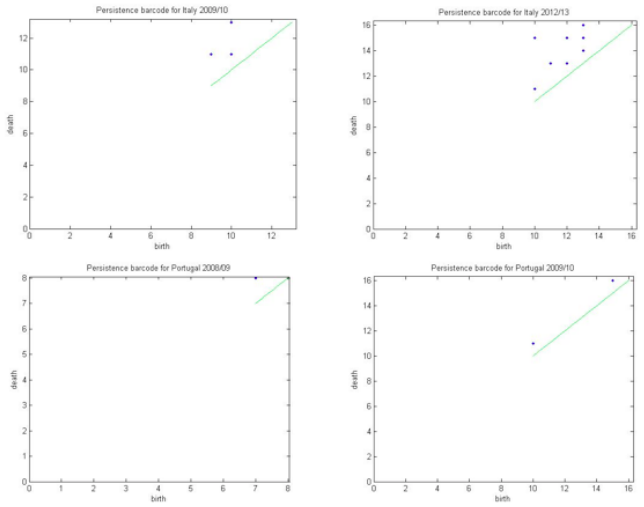


Fig. 8. Comparing the ILI seasons using persistence diagrams for dimension 1 for: Italy 2009/10 (on the upper left), Italy 2012/13 (on the upper right), Portugal 2008/09 (on the lower left), and Portugal 2009/10 (on the lower right), identified as particular cases in Figure 7.

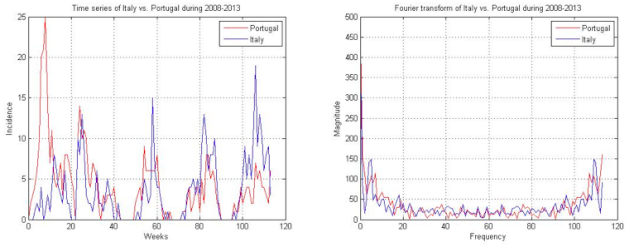


Fig. 9. Comparing the ILI seasons of Portugal and Italy during 2008-2013: the time series (on the left); the Fourier transform (on the right).

within pairs of numbers representing birth and death times. We have computed a persistence diagram in Figure 4 for each time series (country, year) embedded in higher dimensions. As shown by the persistence diagrams below, the distinguishable features are seen in dimension 1.

V. CONCLUSION AND FUTURE WORK

The study of Epidemiology is a great source of problems relating to nonlinear systems, large-scale data and development of more accurate models, where TDA can contribute, providing high dimension techniques for medical data analysis. In this study we showed how they could be used to analyze and compare ILI seasons between countries based on the curves of the time series of their ILI incidence. The analyzed Influenzanet data lists the number of active participants and the number of ILI onsets, for three different ILI case definitions. Using the described methods we shall also look at those different ILI case definitions, contributing to a better understanding of the features distinguished by them. The information provided by quantitative methods such as DTW or the Fourier analysis of time series can be combined and complemented by the TDA analysis of that data. Further research considers the analysis of the impact of the TDA analysis for modeling and prediction of the current Influenza season. We can also complement the approach with other machine learning methods to learn metrics that are more appropriate to the input time series data, aiming to grasp a better understanding of the severity of the epidemics both in past and ongoing ILI seasons.

ACKNOWLEDGMENTS

The first and third authors would like to thank the support of Croatian Science Foundation's funding of the project EVOSOFT (UIP-2014-09-7945), and the H2020 project MIDAS (agreement 727721). The second author thanks the support of ARRS ref. TOPREP (ARRS N1-0058). A special thanks to Gabriela Gomes, co-founder of Influenzanet, for introducing the authors to this participatory monitoring system, for making the data available, and for commenting on a previous version of this work. We are also grateful to Ana O. Franco for information complementing the available data, and for revising a previous version of this work.

REFERENCES

[1] U. Bauer (2015). Ripser. *github.com/ripser*.

[2] G. Carlsson (2009). Topology and data. *Bulletin of the American Mathematical Society* **46.2**: 255–308.

[3] J. M. Chan, G. Carlsson and R. Rabadan (2013). Topology of viral evolution. *Proceedings of the National Academy of Sciences* **110.46**: 18566–18571.

[4] P. Dlotko (2014). Persistence Landscapes Toolbox. *math.upenn.edu/dlotko*.

[5] L. Li et al (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine* **7.311**: 311ra174–311ra174.

[6] V. Nanda (2014). Perseus. *sas.upenn.edu/vnanda/perseus*.

[7] M. Nicolau et al (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences* **108.17**: 7265–7270.

[8] D. Paolotti et al (2014). Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. *Clinical Microbiology and Infection* **20.1**: 17–21.

[9] J. A. Perea and J. Harer (2013). Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Comp.Mathematics* **15.3**: 799–838.

[10] J. Pita Costa and P. Škraba (2014). A topological data analysis approach to epidemiology. In *European Conference of Complexity Science 2014*.

[11] J. Pita Costa and P. Škraba (2015). Topological epidemiological data analysis. In *ACML Health 2015*.

[12] J. Pita Costa (2017). Topological data analysis and applications. In *40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2017)*, IEEE: 558–563.

[13] Sander P. van Noort et al (2007). Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe. *Eurosurveillance* **12.7**: E5–6.

[14] Vin de Silva, P. Skraba, and M. Vejdemo-Johansson (2012). Topological analysis of recurrent systems. In *Workshop on Algebraic Topology and Machine Learning, NIPS 2012*.

[15] E. YomTov et al (2015). Estimating the Secondary Attack Rate and Serial Interval of Influenzalike Illnesses using Social Media. *Influenza and other respiratory viruses* **9.4**: 191–199.

Critical spatial clusters for vaccine preventable diseases

Jose Cadena
Biocomplexity Institute
Blacksburg, VA
jcadena@vt.edu

Achla Marathe
Biocomplexity Institute
Blacksburg, VA
amarathe@vt.edu

Anil Vullikanti
Biocomplexity Institute
Blacksburg, VA
vsakumar@vt.edu

ABSTRACT

Despite high vaccination rates for infectious diseases, such as measles, there have been several big disease outbreaks in recent years. This is, in part, due to misinformation about vaccinations in certain sub-populations, and their spatial clustering. Identifying potential clusters, which can result in big outbreaks in the event of reduced vaccination rate, is an important public health challenge. We develop a natural notion of criticality of such clusters, which extends the problems of influence maximization to connectivity constraints. We develop efficient approximation algorithms for finding critical clusters by exploiting the structural properties of the problem in contact networks. We apply our methods to find critical clusters in the state of Minnesota, with significantly higher criticality than those obtained by heuristics used in public health.

KEYWORDS

Criticality, submodularity optimization, epidemic spread

ACM Reference Format:
Jose Cadena, Achla Marathe, and Anil Vullikanti. 2018. Critical spatial clusters for vaccine preventable diseases. In *Proceedings of ACM KDD Conference (KDD'18).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Many childhood diseases, such as measles and Pertussis, are easily preventable by vaccination. Therefore, it is worrisome that fairly large outbreaks of such diseases have occurred in recent years, such as the measles outbreaks in California in 2015 and in Minnesota in 2017—this is despite high vaccination coverage in the US, e.g., $\sim$ 95% for MMR, the measles vaccine. One of the reasons is the emergence of undervaccinated geographical clusters [17], often driven by misperceptions about side effects of vaccines [4]. The typical response by public health agencies is to monitor these clusters, run active information campaigns, and engage community leaders. However, such interventions are very expensive and time consuming. Another issue is that public health departments might not be aware of all such clusters, especially in the early stages. As a policy design question, public health agencies are interested in discovering which regions are "critical" spatial clusters, where a reduction in vaccination rate could cause a big outbreak. Current practices involve broad outreach efforts to communities considered at risk, which might not be efficient if some communities are not so critical. Formalizing the notion of critical clusters can help public health agencies focus their limited resources in areas where impact can be maximized. Our contributions are summarized below.

KDD'18, August 19-23, London, United Kingdom

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

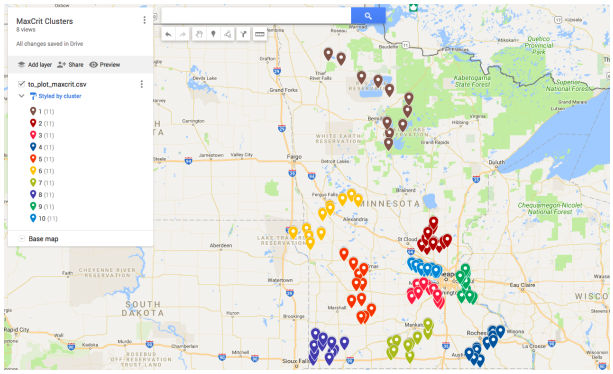


Figure 1: Critical sets in Minnesota discovered using our methods. These are contiguous regions that lead to large outbreaks of measles if not properly vaccinated.

1. Formalizing critical clusters. We define the *criticality* of a subpopulation S as the expected number of *additional* infections that would occur if the individuals in S are not properly immunized. Our focus is on subpopulations located within a bounded spatial cluster. We have different criticality objectives, MaxCrit and ECrit, which capture two distinct public health policy questions: is the source of the infection within the cluster or outside? (Section 4).

Table 1: A summary of our proposed methods

	MaxCrit problem	**ECrit** problem
Motivating policy question	Maximum criticality for source in cluster	Maximum criticality for a fixed source
Characteristics of optimal solutions (Section 3.3.1)	Less connected	Centrally located
Structural property (Section 3.3.2)	Submodular, but not locally modular	Submodular and locally modular
Approximation guarantee	$\Omega(1/k^{1/3})$-approximation (Theorem 3)	$\Omega(1/\log k)$-approximation (Theorem 2)
Empirical observations for MN	Much higher criticality, Most critical cluster is in a rural region	Lower criticality than for MaxCrit, well connected

2. Efficient algorithms for the MaxCrit and ECrit problems. We show that MaxCrit and ECrit are both NP-hard; then, we focus on efficient approximation algorithms for these two problems. Our algorithms exploit structural properties of the objective function

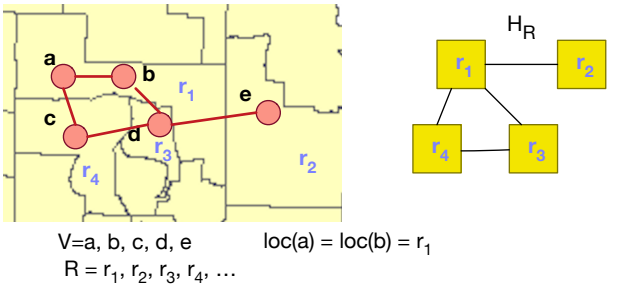


Figure 2: Definitions and notation used in our paper. The 5 red circle nodes (a, b, c, d, e) form a social contact network. Each node resides in a block group r_i, and these block groups from the auxiliary graph $H_{\mathcal{R}}$, where an edge represents that the block groups are neighbors on the map.

and the small world structure of contact networks. MaxCrit and ECrit are instances of submodular function maximization with connectivity constraints, a very challenging problem in combinatorial optimization. However, we show that ECrit has an *approximately modular* structure, which we use to derive a good approximation bound. For the MaxCrit objective, we use the spatial structure to obtain a good approximation factor. Table 1 summarizes these bounds.

3. Application. We evaluate our algorithms on a network model for the state of Minnesota. The sets we discover have very high criticality compared to heuristics commonly considered in the public health community. The critical clusters computed using our algorithms (shown in Figure 1) have meaningful demographic properties from a public health perspective: they typically involve people with lower than average income levels and age (Section 5).

4. Connections with the Influence Maximization problem. We show that criticality is related to the classical problem of influence maximization, but with one very significant difference: *the set of influencers has to form a connected spatial cluster.* As a result, the standard greedy algorithm cannot be used for finding critical clusters. We are not aware of any efficient algorithms with good approximation bounds. The closest is an $\Omega(1/\sqrt{k})$-approximation, which can be obtained by using the algorithm of [15] for submodular function maximization with a connectivity constraint. However, we are able to obtain significantly improved bounds by exploiting the structure of our problems, as summarized in Table 1.

2 PRELIMINARIES

Let V denote a population, and let $G = (V, E)$ be a contact graph on which a disease can spread. That is, a person $v \in V$ (referred to as a "node", henceforth) can propagate the disease to its neighbors. In the social contact network datasets that we consider (Section 5.1.1), each person v is associated with a geographical location, denoted by $\text{loc}(v)$; we will consider such locations at the resolution of census block groups. Let $\mathcal{R}$ denote the geographical region where the nodes V are located—for example, the state of Minnesota—and let $\mathcal{R} = \{r_1, \ldots, r_N\}$ be a decomposition of $\mathcal{R}$ into census block groups. For a block group $r_i \in \mathcal{R}$, let $V(r_i)$ denote the set of nodes v with $\text{loc}(v) \in r_i$. For a subset of block groups $R \subset \mathcal{R}$, let $V(R) = \cup_{r_i \in R} V(r_i)$ be the set of nodes located within it. We consider a graph $H_{\mathcal{R}} = (\mathcal{R}, E_{\mathcal{R}})$ on the set $\mathcal{R}$ of block groups, where two block groups are connected if they are geographically contiguous, i.e., they are adjacent on a map. These definitions are illustrated in Figure 2. Let $\text{Conn}(\mathcal{R})$ denote subsets $R \subset \mathcal{R}$ that are spatially connected in the block group graph $H_{\mathcal{R}}$.

Table 2: Summary of the notation used in the paper

Notation	Description
$G = (V, E)$	Contact graph on set V of individuals
$\text{loc}(v)$	Geographical location of node v
$\mathcal{R} = \{r_1, \ldots, r_N\}$	Geographical region where the nodes V are located—e.g., Minnesota—partitioned into block groups r_i
$V(r_i)$	Set of nodes of G with $\text{loc}(v) = r_i$
$V(R)$	$\cup_{r_i \in R} V(r_i)$
$H_{\mathcal{R}} = (\mathcal{R}, E_{\mathcal{R}})$	Network on $\mathcal{R}$, with adjacent block groups connected by an edge. Sometimes referred to as the "Auxiliary network"
$\text{Conn}(\mathcal{R})$	Set of $R \subset \mathcal{R}$ which are spatially connected in $H_{\mathcal{R}}$
S, E, I, R	States in the disease model
γ	Average region-wide vaccination rate
$\mathbf{x}$	Vaccination vector, with x_i denoting the probability that node i is vaccinated
$\mathbf{x}^R$	Vaccination vector, with nodes $V(R)$ undervaccinated, where $R \in \text{Conn}(\mathcal{R})$
Src_A, Src	Denotes the event that the source of the infection is from a set $A \subset \mathcal{R}$. Src is used when $A = \mathcal{R}$
$\#\text{inf}(\mathbf{x}, \text{Src}_A)$	Expected number of infections for vaccination vector $\mathbf{x}$ and source being Src_A
$\text{crit}(R, \mathbf{x}, \text{Src}_A)$	Criticality of $R \in \text{Conn}(\mathcal{R})$: expected number of additional infections that occur if R is not vaccinated
$\text{MaxCrit}(R)$, $\text{MaxCrit}(R, \mathbf{x})$	The objective value of MaxCrit for region $R \in \text{Conn}(\mathcal{R})$ in an instance $(G, H_{\mathcal{R}}, k)$
$\text{ECrit}(R)$, $\text{ECrit}(R, \mathbf{x})$	The objective value of ECrit for region $R \in \text{Conn}(\mathcal{R})$ in an instance $(G, H_{\mathcal{R}}, k)$

We will use an SEIR model for diseases like measles [3], where a node is in one of *four* states: Susceptible (S), Exposed (E), Infected (I) and Recovered/Removed (R). Measles is highly contagious, and an infected node v spreads the disease to each susceptible unvaccinated neighbor $u \in N(v)$ with high probability. Sometimes, we assume a transmission probability of 1, but all our results extend to the more general case. If v is vaccinated, it does not get infected. We assume the vaccine has 100% efficacy, which is not true in practice, but this is not crucial for our methodology.

Let γ denote the average region-wide vaccination rate—~ 0.97 in Minnesota. Let $\mathbf{x}$ be a *vaccination vector*: $x_i \in [0, 1]$ denotes the probability that node i is vaccinated (so $x_i = \gamma$, by default). Let Src_A denote the source of the infection: this could be one or a small number of nodes from a region $A \subset \mathcal{R}$, which initially get infected (e.g., by contact outside $\mathcal{R}$). We will drop the subscript if $A = \mathcal{R}$. Let $\#\text{inf}(\mathbf{x}, \text{Src}_A)$ denote the expected number of infections for the intervention $\mathbf{x}$, when the initial infection is at Src_A. When the initial conditions are clear from the context, we denote this by $\#\text{inf}(\mathbf{x})$.

2.1 Criticality and Problem Formulations

For a vaccination vector $\mathbf{x}$, let $\mathbf{x}^S$ denote the corresponding intervention where a subset $S \subset V$ of nodes is undervaccinated, and the remaining nodes are vaccinated with the same probability as in $\mathbf{x}$; that is $\mathbf{x}^S_i = \mathbf{x}_i$ for $i \notin S$ and $\mathbf{x}^S_i = \gamma'$ for $i \in S$, where γ' is much lower than γ, the region-wide vaccination rate. For simplicity, we sometimes consider $\gamma' = 0$.

We define the **criticality** of a set $S \subset V$ as $\text{crit}(S, \mathbf{x}, \text{Src}_A) = \#\text{inf}(\mathbf{x}^S, \text{Src}_A) - \#\text{inf}(\mathbf{x}, \text{Src}_A)$, which is the *expected number of additional infections that occur if S is not vaccinated* (with respect to

any specific initial conditions Src_A). Our focus is on finding spatial clusters of high criticality. Specifically, we will focus on $S = V(R)$ for a connected region $R \in \text{Conn}(\mathcal{R})$. We denote this by

$$\text{crit}(R, \mathbf{x}, \text{Src}_A) = \#\text{inf}(\mathbf{x}^R, \text{Src}_A) - \#\text{inf}(\mathbf{x}, \text{Src}_A),$$

which is the expected number of extra infections that might be caused if the nodes in the connected region R are under-vaccinated.

We focus on finding "small" connected regions, since this can lead to an actionable policy for public health agencies. We model this by adding a constraint $|R| \leq k$, where k is a parameter that can be tuned based on the available resources of a public health agency.

We propose two problems that model two different kinds of initial conditions of interest from a public health perspective. The first problem models the following question: *for any specific initial condition (e.g., Src denotes kids in an elementary school), what is the most critical set?*

Problem 1 (k-ECrit$(G, H_{\mathcal{R}}, k)$). *Given an instance $(G, H_{\mathcal{R}}, k)$, find a connected region $R \in Conn(\mathcal{R})$ of size at most k that maximizes criticality:*

$$R = argmax_{R' \in Conn(\mathcal{R}), |R'| \leq k} crit(R', \mathbf{x}, Src)$$

For convenience, we will sometimes also use ECrit$(R, \mathbf{x}, \text{Src})$ or ECrit(R) to denote crit$(R, \mathbf{x}, \text{Src})$, the objective value of ECrit for region R in an instance $(G, H_{\mathcal{R}}, k)$. The second problem models the following question: *what is the most critical cluster if the infection source is the worst possible, which will happen if the infection starts within the undervaccinated cluster itself?* This is formalized as

Problem 2 (k-MaxCrit$(G, H_{\mathcal{R}}, k)$). *Given an instance $(G, H_{\mathcal{R}}, k)$, find a connected region $R \in Conn(\mathcal{R})$ of size at most k that maximizes criticality over all choices of source:*

$$R = argmax_{R' \in Conn(\mathcal{R}), |R'| \leq k, Src_{R'}} crit(R', \mathbf{x}, Src'_R)$$

In other words, the k-MaxCrit problem involves maximizing over all possible choices of the sources $\text{Src}_{R'}$ in the cluster R'. As before, we will use MaxCrit$(R, \mathbf{x}, \text{Src})$ or MaxCrit(R) to denote the objective value of an instance of the problem.

3 KEY PROPERTIES OF CRITICALITY

We start by describing connections between the proposed MaxCrit and ECrit objectives, and influence maximization, which will have implications on the computational complexity. We also prove structural properties of these objectives, later used in our algorithms.

3.1 Complexity and connections with Influence Maximization

In the Influence Maximization (InfMax) problem [12], we are given a directed graph $G = (V, E)$ and edge weights $p(u, v) \in [0, 1]$ indicating the probability that node u influences node v. The *Independent Cascade* model is a special case of the SEIR model, where each node is infectious for exactly one time step. The goal is to find a set $S \subset V$ of k *seed nodes* to infect, such that the expected number of influenced nodes or *spread*, $\sigma(S)$, is maximized. There has been a lot of work on the InfMax problem since its introduction by [12]. An instance of InfMax consists of a single contact graph G, whereas instances of the MaxCrit and ECrit problems consist of the contact graph G, a partition of the nodes of G into regions, $\mathcal{R}$, and an auxiliary graph $H_{\mathcal{R}}$ that captures connectivity among $\mathcal{R}$.

3.1.1 NP-hardness. InfMax can be reduced to MaxCrit and ECrit, which implies their NP-hardness. The proof is by constructing a suitable auxiliary graph $H_{\mathcal{R}}$, a vaccination vector $\mathbf{x}$, and a source Src. This is summarized in Theorem 1, whose proof is presented in the Appendix in the full version of this paper [1].

Theorem 1. *MaxCrit and ECrit are NP-hard.*

3.1.2 Impact of connectivity requirement. The connectivity constraint has a strong effect on the solution of MaxCrit and ECrit. In particular, a solution computed for InfMax using the greedy algorithm of [12] can be arbitrarily suboptimal for the problems we propose. Informally, this follows from the property of InfMax that it is better to choose the set of seeds to be located far apart, so that their combined influence is maximized.

Observation 1. *There exists a family of instances $(G, H_{\mathcal{R}}, k)$ for which the optimum solution S^* to MaxCrit satisfies MaxCrit$(S^*) = O(\frac{1}{k}\text{InfMax}(\hat{S}))$, where $\hat{S}$ is the optimum solution to the InfMax version for this instance, without any connectivity requirements.*

3.2 Submodularity of MaxCrit and ECrit

A set function $f : 2^V \rightarrow \mathbb{R}$ is said to be *submodular* if it satisfies the diminishing returns property: for any $T \subset S \subset V$ and $x \in V \setminus S$, we have that $f(T \cup x) - f(T) \geq f(S \cup x) - f(S)$. We have the following result:

Lemma 3.1. *MaxCrit and ECrit are submodular.*

The proof is presented in the full version, but the argument is similar to the submodularity proof for the InfMax problem.

3.3 Differences between MaxCrit and ECrit

While these two problems model related public health problems, they have some significant differences, both in terms of the structure of the optimum solutions and a locality property, which is useful in designing efficient algorithms.

3.3.1 Difference in the structure of optimal solutions. The solution structure for both problems can be very different in the worst case. Consider a contact graph G split into regions $\mathcal{R} = \{r_1, \ldots, r_N\}$. For each r_i, we assume $V(r_i)$ has n nodes. The auxiliary graph $H_{\mathcal{R}}$ consists of two disjoint sets: graph H_1 induced by $r_1, \ldots, r_{N-k'}$, and graph H_2 induced by $r_{N-k'+1}, \ldots, r_N$. For each $i \leq N - k'$, the graph $G[V(r_i)]$ is a connected component. The graph H_1 forms a chain, with $V(r_1)$ having an edge to $V(r_2)$, $V(r_2)$ having edges to $V(r_1)$ and $V(r_3)$, etc. We have Src to be a node $s \in V(r_{n'})$, where $n' = \lfloor (N - k')/2 \rfloor$. The graph $G[H_2]$ restricted to H_2 is fully connected, but it is disconnected from H_1; we also choose it so that $G[H_2]$ has more nodes than H_1. Then, an optimum solution to the ECrit problem with Src will be the cluster of k block groups centered at $r_{n'}$, with criticality of $O(kn + 2n/\gamma)$, by considering a percolation process on a chain. If we choose $k' > k + 2/\gamma$, the optimal solution to the MaxCrit problem on this instance will be a cluster of k block groups from H_2, since the fully connected structure in H_2 will lead to a larger outbreak.

3.3.2 Local modularity property. ECrit has a local modularity structure, which is motivated by [14]. Specifically, for a set $A_1 \cup A_2 \cup \ldots A_r$ of disjoint and roughly similar sized clusters, ECrit$(A_1 \cup A_2 \cup$

$\ldots A_r) \geq c \cdot \sum_i \text{ECrit}(A_i)$, for a constant $c < 1$. This is different in form from the notion of (r, δ)-local function of [14]. A function $F(\cdot)$ is (r, δ)-local if $F(A_1 \cup A_2) \geq F(A_1) + \delta F(A_2)$ for two sets A_1 and A_2, which are distance r away. It is not clear that ECrit satisfies such property, but the specific kind of property it satisfies is sufficient for using the subsequent technique of [14]. In contrast, we show that the MaxCrit is not locally modular.

We assume our contact graph is a "small world" network, following the model of [13] in which nodes have local connections to nearby nodes and a small number of long range connections. A node u has a long range connection to node v with probability proportional to $\frac{1}{d_{uv}^{\alpha}}$, where α is the "power law exponent", typically $\alpha > 2$. We will consider a set of clusters $A_1, \ldots, A_r$, where A_i has size n_i. We assume all clusters have roughly similar size, so $n_i \leq n_j \beta$ for a constant β. We assume the clusters are small, specifically $n_i \leq \sqrt{n}$. For the analysis below, we assume the disease is highly contagious and there is enough local connectivity within each cluster. Therefore, if nodes in A_i are not vaccinated, and some node $v \in A_i$ gets infected (from outside the cluster), the entire cluster will get infected. We also assume the clusters A_i are fairly localized, so that we can consider d_{ij} to be the distance from the centroid of A_i to that of A_j. For simplicity of the analysis, we will assume that in the small world network model for H, the probability that a node u in A_i connects to a node v in A_j is proportional to $\frac{1}{d_{ij}^{\alpha}}$. The following Lemma—proven in the full version—shows the local modularity of ECrit.

Lemma 3.2. *Let $A_1, \ldots, A_r$ be disjoint clusters, with the model and notation as described above. Then,*

$$\textit{ECrit}(A_1 \cup A_1 \cup \ldots A_r) \geq \frac{1}{1 + 3\gamma\beta/(\alpha - 1)} \Big(\textit{ECrit}(A_1) + \ldots + \textit{ECrit}(A_r) \Big)$$

In contrast, MaxCrit does not satisfy the property from Lemma 3.2. Consider a setting where each block group induces a clique, which is disjoint from all other block groups. Then, $\text{MaxCrit}(A_1) \propto \max_{b \in A_1} |V(b)|$ is proportional to the largest block group in the set A_1. Similarly, we have $\text{MaxCrit}(A_2) \propto \max_{b \in A_2} |V(b)|$. For disjoint sets A_1, A_2, we have

$$\text{MaxCrit}(A_1 \cup A_2) = \max_{b \in A_1 \cup A_2} |V(b)| < \text{MaxCrit}(A_1) + \text{MaxCrit}(A_2)$$

4 PROPOSED METHODS

4.1 Algorithm ApproxECrit

Our algorithm ApproxECrit uses the locality property from Lemma 3.2 and builds on the approach of Krause et al. [14] and Borgs et al. [5]. Algorithm 1 gives a pseudocode description, and we give the intuitive ideas below.

(1) **Padded decompositions.** This is a partition of the graph $H_{\mathcal{R}}$ into clusters $C_1, \ldots C_\ell$, each of diameter at most $12r$. If a node v and all nodes at distance at most r of v are in the same cluster, we say that v is r-padded. After clustering, all the nodes that are not r-padded are removed; this occurs with probability 1/2 for each node, and the best solution S of size k after removal has objective value $F(S) \geq \frac{1}{2} F(S^*)$, where S^* is the optimal subgraph of size k.

(2) **Greedy solution in the clusters.** The purpose of the padded decomposition was to partition the graph into small clusters where we can ignore the connectivity cost [14]. For each cluster, we now run the greedy algorithm for submodularity maximization to obtain an ordering of the nodes; the first j nodes in this ordering are approximately the most informative nodes in the cluster. We implement Algorithm 2, a modified version of the algorithm in [5] to account for the fact that we want a graph that is connected in the auxiliary graph, but with the epidemic process occurring in the social contact network. The greedy algorithm degrades the quality of the optimal solution by a factor of at most $(1 - 1/\epsilon)$.

(3) **Running Quota Steiner Tree on $\mathcal{R}$.** Finally, we compute wt_r for each $r \in \mathcal{R}$, and then compute a quota Steiner tree T' of size k, which maximizes $\sum_{r \in T'} \text{wt}_r$. The subroutine k-MaxST uses the fixed parameter algorithm of [6] to find an optimal solution, as described in Section 4.2.1.

Algorithm 1 ApproxECrit(G, $H_{\mathcal{R}}$, k, Src).

1: Partition $H_{\mathcal{R}}$ into clusters $C_1, \ldots C_\ell$, each of diameter at most $12r$, using the method of [14] (referred to as a *Padded decomposition*)
2: For each cluster $C_i = \{r_{i1}, \ldots, r_{ij}\}$, let $(r_{ia_1}, r_{ia_2}, \ldots, r_{ia_j}) =$ Greedy(C_i, j, Src), be an ordering of block groups obtained by running Greedy without connectivity constraints
3: Construct a connected graph G' on the nodes $\{r_{ia_1} : i = 1, \ldots, \ell\}$ with an edge (r_{ia_1}, r_{ja_1}) having weight equal to the shortest path length in $H_{\mathcal{R}}$. Run the Budgeted Steiner Tree algorithm of [11] to find a tree T with k nodes and maximum total criticality
4: **for** $r \in H_{\mathcal{R}}$ **do**
5: Let $\text{wt}_r = \text{crit}(r)$
6: **end for**
7: Let $T' = k - \text{MaxST}(H_{\mathcal{R}}, \text{wt}, k)$ using the algorithm of [6]
8: return $\max\{\text{ECrit}(T), \text{ECrit}(T')\}$

Algorithm 2 Greedy(C, j, Src).

1: $\mathcal{S} = \phi$, $L = cj|E| \log |V|$, for a constant c
2: $\ell = 0$
3: **while** $\ell < L$ **do**
4: Pick random subgraph G' of G with (1) edges sampled based on disease transmission probability, (2) node $v \in V(C)$ sampled with probability $1 - \gamma'$, (3) $v \notin V(C)$ sampled with probability $1 - \gamma$
5: $\ell = \ell + |E(G')|$
6: Let C_i be the set of components reachable from Src in G'
7: $\mathcal{S} = \mathcal{S} \cup \{C_i\}$
8: **end while**
9: Initialize $X = \phi$
10: For each $r \in C$, define $deg(r, \mathcal{S})$ to be the number of sets $C_i \in \mathcal{S}$ that contain some node in $V(r)$
11: **for** $i = 1$ to j **do**
12: Append $r = \text{argmax}_{r'} deg(r', \mathcal{S})$ to X
13: Remove all sets C_i hit by $V(r)$ from $\mathcal{S}$ and update all $deg(\cdot)$
14: **end for**

Theorem 2. *Let S^* denote an optimal solution to an instance of the ECrit$(G, H_{\mathcal{R}}, k, \mathbf{x}, Src_A)$ problem. Let S be the cluster returned by ApproxECrit. If $H_{\mathcal{R}}$ forms a small world network, and the sizes of all block groups in $\mathcal{R}$ are within a constant factor of each other, then S has $O(k)$ nodes and ECrit$(S, \mathbf{x}, Src_A) \geq \Omega(\frac{1}{1 + 3c'\gamma\beta/(\alpha-1)})$ECrit$(S^*, \mathbf{x}, Src_A)$, where γ, β and α are as defined in Lemma 3.2. The worst case running time of ApproxECrit is $O(|\mathcal{R}||E|k(2e)^k)$.*

4.2 Algorithm ApproxMaxCrit

Algorithm 3 ApproxMaxCrit(G, $H_{\mathcal{R}}$, k).

1: **for** $r \in H_{\mathcal{R}}$ **do**
2: Let C_r be the set of block groups within distance $B = O(k^{2/3})$ of r in $H_{\mathcal{R}}$. Construct graph $H_{\mathcal{R}}[C]$ induced by the block groups in C
3: Run Greedy(C, B) with the following modification: the source in the sampling step is picked from $V(C)$ randomly in each iteration. Let $r_1, r_2, \ldots, r_B$ be the block groups which are picked
4: Construct a minimum Steiner tree T_r of $r_1, \ldots, r_B$
5: **end for**
6: **for** $r \in H_{\mathcal{R}}$ **do**
7: Let $\text{wt}_r = \text{crit}(r)$
8: **end for**
9: Let $T' = k - \text{MaxST}(H_{\mathcal{R}}, \text{wt}, k)$ using the algorithm of [6]
10: return $\max\{\max_r \text{MaxCrit}(T_r), \text{MaxCrit}(T')\}$

Algorithm ApproxMaxCrit uses ideas from [15], who consider the problem of connected submodular function maximization. Theorem 3 gives a significantly better approximation bound with better running time than [15] by exploiting the spatial properties of our problem. As in the case of ApproxECrit, we also consider a quota Steiner tree and take the best of the two solutions.

Theorem 3. *For an instance $(G, H_{\mathcal{R}}, k)$, let $\hat{S}$ be the solution returned by ApproxMaxCrit. Let S^* be the optimum solution for this instance. If the aspect ratio of the bounding box containing $\mathcal{R}$ and each block group is constant, MaxCrit$(\hat{S}) \geq \Omega(\frac{1}{k^{1/3}})$MaxCrit$(S^*)$. The worst case running time is $O(|\mathcal{R}|k^{2/3} + |\mathcal{R}||E|k(2e)^k)$.*

Proof. (Sketch) For simplicity, assume each block group is a square; the arguments extend easily with a constant factor increase in the approximation bounds, since the aspect ratios are constant. Our proof is in two parts: (1) for any r, the Steiner tree T_r has at most k nodes, (2) there is a set of $O(k^{1/3})$ trees $T_{r'_1}, \ldots, T_{r'_s}$, such that they together cover S^*. We first argue that the theorem follows from these two statements. Statement (1) above implies that each T_r is a feasible solution to k-MaxCrit, since T_{r_i} is a connected subgraph in $H_{\mathcal{R}}$. Statement (2) implies $\sum_i \text{MaxCrit}(T_{r'_i}) \geq \text{MaxCrit}(S^*)$, by submodularity. Thus, there exists a node r_i such that $\text{MaxCrit}(T_{r'_i}) \geq \Omega(1/k^{1/3})\text{MaxCrit}(S^*)$, and the theorem follows.

We now prove statement (1). We consider any node r in $H_{\mathcal{R}}$. First, observe that a set of $O(k^{2/3})$ square subgraphs, each of side $O(k^{1/3})$ covers $H_{\mathcal{R}}$; let these be $y_1, \ldots, y_s$. Next, there exists a tree T' of length $O(k^{2/3} \cdot k^{1/3}) = O(k)$ that connects the centers of all the squares y_i. Then, T' can be augmented with additional paths to connect all the nodes $r_1, \ldots, r_B$, with only a constant factor increase in the number of nodes. This follows because each r_i is within some square y_j of size $O(k^{1/3}) \times O(k^{1/3})$, so that it can be connected to T' with a path of length at most $O(k^{1/3})$. Since $B = O(k^{2/3})$, tree T_r connects all the r_j's with a total length of $O(k)$.

Finally, we prove statement (2). Consider a tree T^* spanning S^*. We find the trees $T_{r'_1}, \ldots$ above in an iterative manner. First, pick a leaf r'_1 of T^*, and remove from T^* all the block groups which are within distance $k^{2/3}$ of r'_1, and repeat the process on the residual tree. Each such tree $T_{r'_1}$ covers at least $\Omega(k^{2/3})$ nodes of T^*. Therefore, $O(k^{1/3})$ trees computed in this manner cover T^*. □

4.2.1 Subroutine k-MaxST for the quota Steiner tree problem. Both algorithms ApproxECrit and ApproxMaxCrit involve solving an instance of the quota Steiner tree problem: given a graph $H_{\mathcal{R}}$, a weight wt_r for each $r \in \mathcal{R}$, and a parameter k, the objective is to compute a tree T' in $H_{\mathcal{R}}$ with at most k nodes, such that $\sum_{r \in T'} \text{wt}_r$ is maximized. There are constant factor approximations for this problem [21]. Here, we adapt the randomized fixed-parameter tractable algorithm of Cadena et al. [6] for Prize-Collecting Steiner Tree, which gives an optimal solution with high probability. The algorithm relies in the seminal color-coding technique of Alon et al. [2]. Naively, one could find a solution to k-MaxST by exhaustively checking all the possible $\binom{n}{k}$ subgraphs of k nodes in time $O(n^k)$. The algorithm does a random k-coloring of the nodes of $H_{\mathcal{R}}$, and it only considers maximum weight trees of each size that are "colorful"—this means all the nodes have distinct colors. It can be shown that such colorful solutions can be computed using a dynamic program. Further, the optimal solution is colorful with probability $k!/k^k$, which is large enough for the algorithm to work. Thus, the color coding technique allows us to reduce the search space to $O((2e)^k)$, keeping the computation feasible.

5 EXPERIMENTAL RESULTS

Our experiments focus on the following questions:

(1) **Finding critical clusters.** Can we find highly critical regions with our proposed methods? How do they compare to standard baselines used in public health? (Section 5.2)
(2) **Demographics.** What are the demographic properties of critical clusters? Where are they located? (Section 5.3)
(3) **MaxCrit vs. ECrit** What are the differences and similarities of the clusters discovered under the two proposed problems? (Section 5.4)

5.1 Experimental Setup

5.1.1 Dataset and disease model. A study of epidemics that spread through physical proximity requires social contact networks in which an edge represents an actual physical contact between two people at some location during the day. Such networks are not readily available and cannot be constructed easily because of the difficulty in tracking contacts for a large set of people. This has been recognized as a significant challenge in the public health community, and multiple methods have been developed to construct large scale realistic contact network models by integrating diverse public datasets (e.g., US Census, land use and activity surveys) and commercial data (e.g., from Dunn & BradStreet on location profiles). We use models developed by the approach of [8];[1] see also [9, 19] for network models developed by other public health groups.[2] Multiple such network models were evaluated in a study by the Institute of Medicine [10].

Here, we focus on a population for Minnesota with 5,048,920 individuals in total, which are aggregated into 4,082 census block groups from the 2010 U.S. census. We consider an SEIR type of stochastic model for measles, as described earlier in Section 2. For the MaxCrit formulation, the criticality of a cluster C of block groups

[1] See ndssl.vbi.vt.edu/synthetic-data/download for networks available for download.
[2] Models are available at http://www.epimodels.org/drupal/?q=node/70 and https://www.rti.org/impact/synthpop

is assessed by leaving every individual inside C unvaccinated; everybody else in the population is vaccinated with probability 0.97, which is the statewide vaccination rate. The source Src is picked as a set of three nodes in C. For the ECrit formulation, we focus on the Minneapolis metropolitan area, and pick Src to be a set of 100 children. As before, we assess the criticality of a cluster by leaving its inhabitants unvaccinated, with a 0.97 vaccination rate elsewhere.

5.1.2 Baseline Methods. We compare our algorithms with two heuristics used in epidemiology and a naive random baseline.

(1) Population. Find a cluster of size k with the largest total population. The motivation behind this heuristic is leaving as many people as possible unvaccinated.
(2) Vulnerability. The vulnerability of an individual is the probability that this person will get infected when the disease is left to propagate with no intervention—i.e., $x_v = 0$ for all nodes. This baseline finds a cluster of size k with as large total vulnerability as possible, thus prioritizing individuals who are most likely to get infected.
(3) Random. Find a connected cluster of size k by doing a random walk on the auxiliary graph.

5.2 Optimization power

In Figure 3, we show the criticality obtained by ApproxMaxCrit (top) and ApproxECrit (bottom) compared to the three baseline methods as a function of k. As expected, selecting subgraphs at random performs poorly and results in almost no additional infections compared to the initial disease conditions. Surprisingly, Vulnerability does not perform much better than random, especially on the MaxCrit objective. It is also interesting that the population-based heuristic does not have monotonic improvement with k. For the top plot, even though the subgraph of size 9 has 55,800 inhabitants, the smaller subgraph of size 5 with a population of 34,000 leads to a significantly larger outbreak. Overall, the population-based heuristic has better performance among the baselines, and it even surpasses our algorithm for $k = 5$ in MaxCrit. However, both ApproxMaxCrit and ApproxECrit exhibit notably better performance. For the ECrit objective, the 11-node cluster discovered using our method leads to 8 times more infections than the baselines.

Another important quantity is the probability of having a large outbreak. In Figure 4, we show the distribution of criticality values for each method over 100 simulations of the disease model. For the MaxCrit objective (top), we observe that even the largest outbreaks caused by Vulnerability and Random are much smaller than those of ApproxMaxCrit and the Population baseline. We also note that the population-based clusters have larger variance in criticality and can result in larger outbreaks than those from our algorithm. We observe a similar effect on the ECrit formulation (bottom), where the 9-node Population cluster has extreme cases with more infections than ApproxECrit. This suggests that if the goal for a public health department is to prevent the worst-case scenario, then intervening the most-populated areas is a good heuristic. However, in doing so, one could miss smaller regions that, on average, are likely to infect more people.

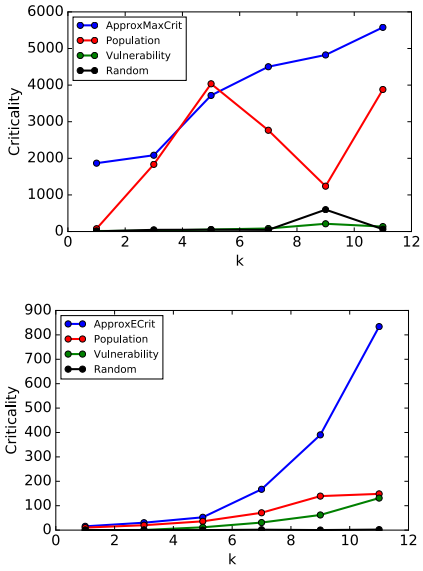


Figure 3: Comparison of algorithms for MaxCrit (top) and ECrit (bottom) as a function of the solution size k

5.3 Critical clusters and demographics

We compare the distribution of age and income in the cluster discovered by ApproxMaxCrit ($k = 11$) to that of the entire state. We aggregate household income into "Low" (below $25,000), "Medium" (between $25,000 and $75,000), and "High" (above $75,000). Ages are binned into "Pre-school" (below 5 years old), "School" (between 5 and 18 years old), "Adult" (between 18 and 70 years old), and "Senior" (above 70 years old). In Figure 5, we see the critical cluster has significantly more households of low income compared to the entire state—19.6% to 34.9%. Similarly, in the discovered cluster, children are over-represented. 26.6% of the population are children in "School" age compared to the national average of 18.7%.

We find critical clusters in different regions over Minnesota. Figure 1 shows the top 10 non-overlapping clusters discovered using ApproxMaxCrit. The most critical cluster—with over 5,000 infections—is located on the rural northern part of the state, spanning the Leech Lake and Red Lake reservations. We note that this cluster results in the largest spread despite having a relatively small population of 14,910 people, compared to clusters in urban regions. For example, the second most critical cluster—north of Minneapolis—has 48,889 inhabitants.

In addition to analyzing the most critical cluster, we look at the top-5 non-overlapping clusters discovered by ApproxMaxCrit. These correspond to different choices of root on the k-MaxST algorithm. In Table 3, we report the total population size, criticality, and percentage of infections to the total population of the cluster—i.e., criticality / population. Note that this latter number could be larger than 1, since there are infections outside the cluster. As we discussed before, the top region leads to a large spread (41% of its population size) despite having less inhabitants than the successive clusters. However, the second cluster follows right after, with virtually the same criticality score, but in a more urban region.

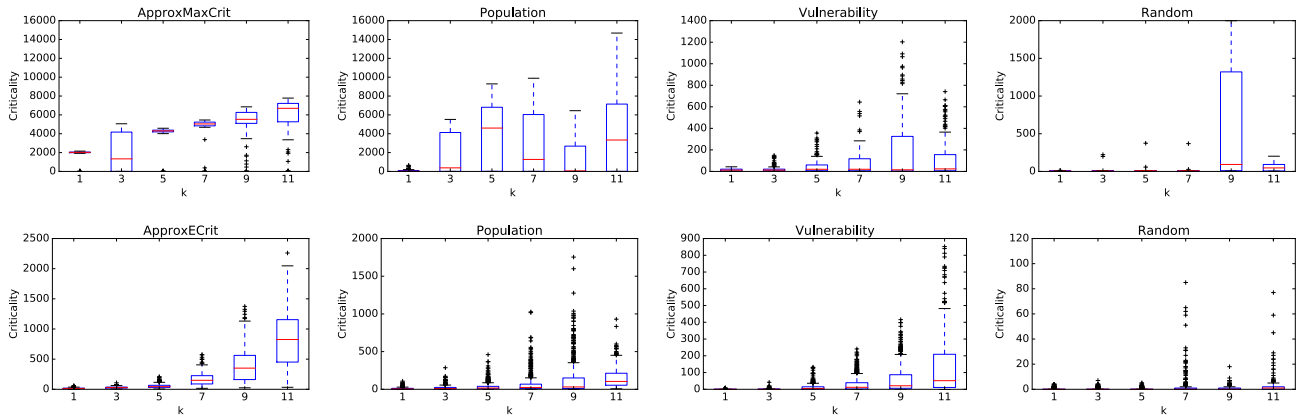


Figure 4: Criticality scores on the MaxCrit objective (top) and ECrit objective (bottom) over 100 runs of the simulation for each method evaluated

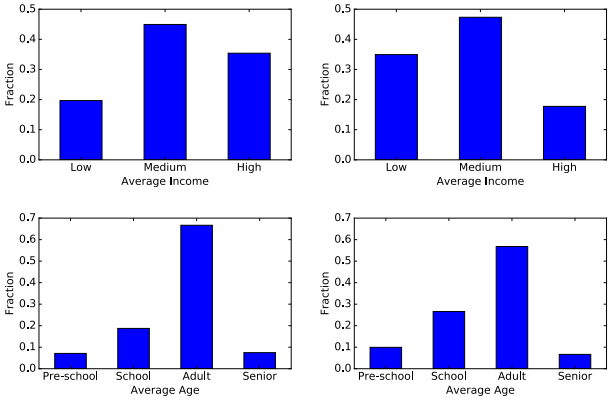


Figure 5: Average income (top) and age (bottom) in the entire state (left) and in the cluster discovered by ApproxMaxCrit. There are more children in school age and lower income households in the discovered critical cluster.

Table 3: Total population and criticality in the top 5 clusters discovered by ApproxMaxCrit

Rank	Population	Criticality	% population
1	14,910	6,138	41.2%
2	48,889	6,093	12.5%
3	23,391	1,388	5.9%
4	15,731	647	4.1%
5	9,936	372	4.7%

For ECrit, we focus on Minneapolis. In Figure 6, we show the most critical clusters for this region. The cluster that produces the largest spread covers the city of Brooklyn Park, which is a "majority-minority" suburb with a large immigrant population.[3] However, we emphasize the need for domain-expert analysis to better interpret and make use of these results. In Table 4, we report the population

[3]https://tinyurl.com/y97k7y2l

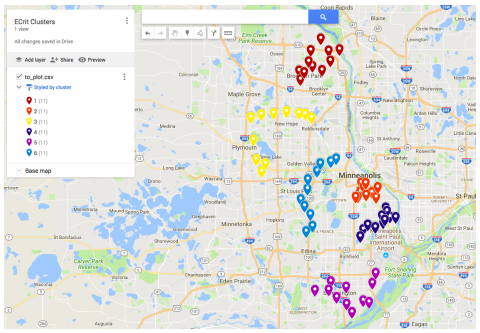


Figure 6: Critical clusters in Minneapolis on the ECrit objective with seeds being children of ages 10 and below.

Table 4: Total population and criticality in the top 5 clusters discovered by ApproxECrit

Rank	Population	Criticality	% population
1	22,273	858	3.9%
2	20,149	58	.3%
3	12,248	40	.3%
4	8,998	14	.2%
5	9,620	12	.1%

and criticality for the 5 most critical clusters. The difference in criticality between the first and second clusters is striking even though their population size is very similar.

5.4 MaxCrit and ECrit

In order to compare clusters from both formulations, we repeat our experiments for the MaxCrit objective on the Minneapolis area instead of the entire state. We find that the clusters discovered with both formulations overlap by a large margin. In Figure 7, we show the MaxCrit clusters in orange circles and the ECrit clusters in blue markers. Not only do the clusters cover the same parts of Minneapolis, but the criticality ranking is the same too. For instance,

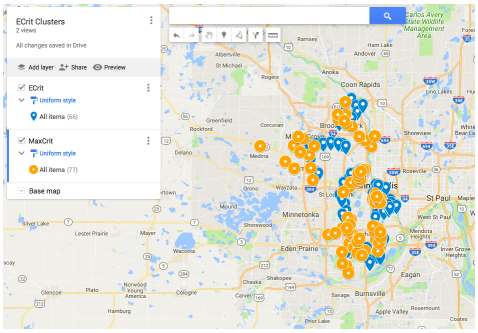


Figure 7: MaxCrit and ECrit clusters in Minneapolis. Both solutions find similar critical clusters.

the most critical cluster using MaxCrit covers Brooklyn Park, just as the ECrit cluster that we discussed in the previous section; this result holds even though the seeds for MaxCrit are chosen from the entire population, whereas we chose children only for ECrit.

6 RELATED WORK

Traditionally, epidemiological models have been differential equation models, which assume very simplistic mixing patterns of the underlying population. In the last decade, a number of research groups have developed agent-based methods using complex network models as a way to handle these issues [8–10, 18, 19]. Such methods have been used for policy analysis by local and national government agencies [10]. Since data for large scale contact networks is not available, we use this paradigm in our work.

All prior work on undervaccinated clusters has been restricted to identification. For instance, Lieu et al. [17] analyze electronic health records among children in 13 counties in Northern California and identify various significant geographic clusters of underimmunization and vaccine refusal, using spatial scan statistics. However, such methods are not directly useful for the policy questions of identifying *critical* clusters, which is our focus here.

There has been a lot of work on different kinds of detection problems related to outbreaks in networks. For instance, Christakis and Fowler [7] use the "friend of random people" approach to monitor a subset of people and infer characteristics of the epicurve for the entire population. Leskovec et al. [16] study the problem of early detection of different kinds of events—e.g., in water networks or social networks. However, these approaches have been focused on either just detecting that some event (e.g., start of an infection) has occurred or the epidemic characteristics for the entire region. Instead, we are interested in finding regions that would lead to a big number of infections if left unvaccinated.

Our work is also related to submodular function maximization with connectivity constraints. This constraint makes the problem much harder than other constraints, such as cardinality or matroid constraints, which can be approximately optimized using a simple greedy procedure [20]. The most relevant work is by Kuo et al. [15], who proposed a $\Omega(1/\sqrt{k})$ approximation algorithm to this problem. We are able to obtain an improved $\Omega(1/k^{1/3})$ approximation for MaxCrit by exploiting the spatial structure in our problem. Finally, Krause et al. [14] propose an approximation algorithm for budgeted submodularity maximization on graphs based on exploiting local structure. Our algorithm for ECrit builds on this approach by exploiting a slightly different type of local modularity bound.

7 CONCLUSIONS

Our work is motivated by public health policy questions of quantifying potential risks of large outbreaks as a result of reducing vaccination rates in a cluster. We formalize two problems, ECrit and MaxCrit, for finding critical clusters for highly contagious diseases that can be prevented by vaccination. These two formulations have different properties and solution structure, and they capture two different policy questions. We show that these problems are variants of the classical influence maximization problem, with an additional connectivity requirement on an auxiliary network, and we design algorithms with rigorous approximation guarantees. Experimental results show that our formulations perform significantly better than heuristics from epidemiology. Such an approach can help public health agencies prioritize response to the challenges of reduced vaccination coverage.

REFERENCES

[1] 2018. Critical spatial clusters for vaccine-preventable diseases. https://tinyurl.com/yc2tw7u7. (2018).
[2] Noga Alon, Raphael Yuster, and Uri Zwick. 1995. Color-coding. *Journal of the ACM (JACM)* (1995).
[3] R.M. Anderson and R.M. May. 1991. *Infectious Diseases of Humans.* Oxford University Press, Oxford.
[4] Jessica E. Atwell et al. 2013. Nonmedical Vaccine Exemptions and Pertussis in California, 2010. *Pediatrics* (2013).
[5] Christian Borgs et al. 2014. Maximizing Social Influence in Nearly Optimal Time. In *Proc. SODA*. 946–957.
[6] Jose Cadena, Feng Chen, and Anil Vullikanti. 2017. Near-Optimal and Practical Algorithms for Graph Scan Statistics. In *SIAM Data Mining (SDM)*.
[7] N.A. Christakis and J.H. Fowler. 2010. Social Network Sensors for Early Detection of Contagious Outbreaks. *PloS one* 5, 9 (2010), e12948.
[8] S. Eubank et al. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature* 429 (2004), 180–184. Issue 6988.
[9] N.M. Ferguson, D.A.T. Cummings, C. Fraser, J.C. Cajka, P.C. Cooley, and D.S. Burke. 2006. Strategies for mitigating an influenza pandemic. *NATURE-LONDON-* 442, 7101 (2006), 448.
[10] M. Halloran et al. 2008. Modeling targeted layered containment of an influenza pandemic in the United States. In *PNAS*. 4639–4644. PMCID:PMC2290797.
[11] D. Johnson, M. Minkoff, and S. Phillips. 2000. The Prize Collecting Steiner Tree Problem: Theory and Practice. In *ACM SODA*.
[12] D. Kempe, J. Kleinberg, and É. Tardos. 2003. Maximizing the spread of influence through a social network. In *KDD*. 137–146.
[13] J. Kleinberg. 2000. The small world phenomenon: An algorithmic perspective. *Proceedings of STOC* (2000).
[14] Andreas Krause et al. 2006. Near-optimal sensor placements: Maximizing information while minimizing communication cost. In *International conference on Information processing in sensor networks*. ACM, 2–10.
[15] Tung-Wei Kuo, Kate Ching-Ju Lin, and Ming-Jer Tsai. 2015. Maximizing Submodular Set Function With Connectivity Constraint: Theory and Application to Networks. *IEEE/ACM Transactions on Networking* 23, 2 (2015), 533–546.
[16] Jure Leskovec et al. 2007. Cost-effective outbreak detection in networks. In *KDD*. 420–429.
[17] Tracy A Lieu et al. 2015. Geographic clusters in underimmunization and vaccine refusal. *Pediatrics* 135, 2 (2015), 280–289.
[18] F. Liu et al. 2015. The role of vaccination coverage, individual behaviors, and the public health response in the control of measles epidemics: an agent-based simulation for California. *BMC Public Health* 15, 1 (01 May 2015), 447.
[19] Ira M. Longini et al. 2005. Containing Pandemic Influenza at the Source. *Science* 309, 5737 (August 2005), 1083–1087.
[20] GL Nemhauser, LA Wolsey, and ML Fisher. 1978. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming* 14, 1 (1978), 265–294.
[21] Ramamurthy Ravi et al. 1996. Spanning trees—short or small. *SIAM Journal on Discrete Mathematics* 9, 2 (1996), 178–200.

Identification of At-Risk Groups for Opioid Addiction Through Web Data Analysis

A work in progress report submitted to epiDAMIK workshop to be held in conjuction with ACM SIGKDD 2018

K. Basu, S. Choudhuri, A. Sen
NetXT Lab, SCIDSE
Arizona State University
{kaustav.basu, schoud13, asen}@asu.edu

A. Majumdar
Galvanize, Inc.
aniket.majumdar@galvanize.com

D. Dey
Dept. of Statistics
University of Connecticut
dipak.dey@uconn.edu

ABSTRACT

The Opioid epidemic that claimed more than 63,600 lives in 2016, was declared as a public health emergency by the US government in October 2017. Although a few health insurance companies and commercial firms have examined this important issue from various available data sources, research findings from analysis of publicly available Opioid related web data is sparse. Accordingly, we have undertaken the important task of *identification of at-risk groups for Opioid addiction* through web data analysis, so that appropriate early intervention measures can be initiated by public health officials. We have collected Opioid incidences data for the states of Connecticut and Ohio for the time period of 2012 - 2018, and we are currently in the process of analyzing such data. In this paper, we present our preliminary findings and outline our plans for further research on this topic.

CCS CONCEPTS

• **Computing methodologies** → *Unsupervised learning*;

KEYWORDS

Opioid Addiction, Web Data, Risk Group Identification

1 INTRODUCTION

Opioids are drugs, prescribed by health professionals to relieve patients from pain. Unfortunately, these drugs often lead to addiction. This addiction has emerged as a full blown epidemic in the United States. In the last few years, there has been an alarming increase in Opioid related deaths, resulting in the loss of 63,600 lives in 2016 alone. In October 2017, the epidemic was declared as a public health emergency by the US government. Although a few health insurance companies and commercial firms have examined this important issue from various available data sources, research findings from analysis of publicly available Opioid related web data is sparse. Accordingly, we have undertaken the important task of *identification of at-risk groups for Opioid addiction* through web data analysis, so that appropriate early intervention measures can be initiated by public health officials. We have collected Opioid incidences data for the states of Connecticut and Ohio for the time period of 2012 - 2018, from various federal, state and local government databases, and we are currently in the process of analyzing such data. In this paper, we present our preliminary findings and lay out our plans for further research on this topic, which also includes finding the *pathways to Opioid addiction*. From the available literature, it appears that a major pathway to Opioid addiction is through drugs prescribed by medical professionals, to alleviate chronic pain of their patients. Our goal is to find out if this is the *only* pathway to Opioid addiction, or there exists other pathways, such as *peer pressure*, which is recognized as a pathway to alcohol and other non-Opioid drug addiction.

For our analysis, we have collected Opioid incidences data for the states of Connecticut and Ohio for the period 2012 - 2018. In particular, we have collected, (i) the Accidental Drug Related Deaths [1] dataset, for the state of Connecticut for 2012-2017, (ii) the USDA Economic Research Service dataset [2] for 2016-2017, (iii) the Centers for Medicare and Medicaid Services (CMMS) [3] dataset of 24 million Opioid related prescriptions written by 1 million unique prescribers in U.S. during 2014, (iv) A subset of CMMS dataset with 25,000 unique prescribers available on [4] and used by IBM researchers [5], (v) the Cincinnati Heroin Overdose dataset for 2015 - 2018 [6], and, (vi) Cincinnati neighborhood dataset for median income, median age and educational distribution [7, 8]. Brief descriptions of these datasets are provided in Section. 3.

In our effort to identify at-risk groups for Opioid addiction, we focus on the eight counties in Connecticut. Based on available data, our goal is to identify at-risk groups by taking *six* different factors - *location, race, gender, age, income and education level* into account. For our study, location is identified by one of the eight counties of Connecticut. We divide race into five categories (White, African American, Asian, Hispanic/Latino and Others), gender into three categories (Male, Female and Others) and age into four categories (Below 20, 20-39, 40-59 and 60 and above). Income level is divided into four categories (Less than $30k, $30k-$60k, $60k-$100k and above $100K), and education level is also divided into four categories (less than high school, graduated high school, some college, college graduate). Ideally, at the end of our analysis, we expect to identify

epiDAMIK @KDD '18, August 20, 2018, London, United Kingdom

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

at-risk groups at the following level of granularity and be able to make statements such as: *"White Male residents of Fairfield County in the age group of 20-39, within income bracket $30k-$60k, and educational level high school graduate"*, are the highest risk group in Connecticut.

In addition to identification of at-risk groups for Opioid addiction and pathways to addiction, we also examined the important contributing factors of the Opioid epidemic by attempting to answer the following questions,

- Q1: Is there a correlation between the prescribers, prescriptions and Opioid related deaths in U.S. states?
- Q2: Which prescribers are likely to prescribe more than 10 Opioid related prescriptions in a year?
- Q3: Is there a correlation between the income level, age, and the educational level and the Opioid related incidences, in a neighborhood?

From our analysis of historical data, we plan to identify the *characteristics* of risk groups and utilize it to develop a model to predict emerging risk groups in a state. In particular, we plan to develop a unsupervised learning techniques such as, clustering, for risk analysis of specific demographic groups. Using such a model, we plan to forecast the spatial effect (urban or rural) of the emerging risk groups.

In our ongoing effort, so far, we have analyzed (i) county level data of Connecticut, to rank various at-risk groups, according to their vulnerability to addiction, (ii) neighborhood level data of a specific city in Ohio (Cincinnati) to identify the impact of income level, age and educational level, on Opioid related incidences, (iii) national level data to identify the role of the individuals prescribing Opioid drugs on the spread of the epidemic.

2 RELATED WORK

In [9], the authors develop a model to identify patients at risk for prescription Opioid abuse, their dependence and misuse, using drug claims data. For exploration of pathways to prevention, the authors in [10] studied the effectiveness and risks of long term Opioid therapy for chronic pain. Machine learning techniques for surveillance of drug overdose was studied in [11]. Illicit sales of Opioid drugs, such as Fentanyl, through Twitter, was studied in [12]. Chary et. al. in [13] also analyzed Twitter data with a goal of identifying the location of the Opioid related Tweet. Data Science researchers from IBM Research and experts from IBM Watson Health have recently [14] undertaken studies in this domain. Their effort is directed towards the analysis of the relationship between factors surrounding an initial opioid prescription, and a subsequent diagnosis of addiction. The goal of this research is to identify causal factors that lead to addiction diagnosis, taking into account all the variables associated with the initial prescription, such as opioid class, quantity, and related medical procedures and diagnoses.

3 DATASETS FOR ANALYSIS

In order to identify the various at-risk groups, we first collected data from multiple sources and then munged the collected data to create additional datasets. The details of our data collection and data munging are provided in Sections 3.1 and 3.2.

3.1 Data Collection

Our collected data comprises of datasets DS1-DS7. In the following we describe each one of them.

3.1.1 DS1: It corresponds to the Accidental Drug Related Deaths 2012-2017, for the state of Connecticut [1]. This dataset comprises of 4083 unique incidences, across the state of Connecticut, during 2012-2017. Each record in DS1 includes Incidence Date, Gender, Race, Age, Residence/Death City/County of the individual involved, Location of Incidence in latitude/longitude and also the environment - hospital, residence, etc. Moreover, the records contains information related to description of injury, including drug use, substance abuse, multiple medications, etc. In addition, it provides the information about the immediate cause and specific type of Opioid and/or other drugs involved in the incidence.

3.1.2 DS2: The USDA Economic Research Service dataset [2] contains information related to poverty, population, employment/unemployment rates with median household income, and education, for the entire United States for the years of 2016 and 2017. As our focus is in the state of Connecticut, we extract county level information regarding poverty levels to educational levels, for Connecticut, from [2]. We refer to this extracted dataset as DS2.

3.1.3 DS3: The United States Census Bureau American Fact Finder dataset [15] contains information related to demographics, economics, education, etc. for the entire country. We extracted information pertaining to the state of Connecticut, by county, and refer to this extracted dataset as DS3.

3.1.4 DS4: It is the U.S. Opiate Prescriptions/Overdoses dataset available on [4]. This dataset comprises of 25000 unique prescribers, across the U.S., and the prescriptions written by them in 2014. This is a subset of the dataset maintained by the Centers for Medicare and Medicaid Services [3], that contains almost 24 million Opioid related prescriptions, written by 1 million unique health professionals (prescribers), in the U.S in 2014. Each record in DS4 includes *National Provider Identifier number, provider state, gender, credentials and the number of Opioid related drugs prescribed (among the set of 250 different drugs) by the provider.* In addition, it provides the information whether or not the provider prescribed more or less than 10 Opioid related prescriptions in 2014. It may be noted that determination of whether or not a prescriber has prescribed more than 10 prescriptions in 2014, is not done by summing up the number of drugs prescribed by the provider, as multiple drugs may be prescribed on a single prescription.

3.1.5 DS5: This dataset is also collected from [4]. It contains the population in each of the 50 states and also Opioid related deaths in that state.

3.1.6 DS6: It is the Cincinnati Heroin Overdose dataset available on [6]. This dataset is a subset of the Emergency Medical Services (EMS) dataset, where each record contains detailed information regarding an incident, such as location, time, EMS response type, neighborhood, and others, that required an EMS dispatch. This dataset contains information related to Heroin incidences from July 2015 to present time. As of April 18, 2018, there were 5568 such incidences. DS6 is a subset of EMS dataset in the sense that it contains

information only regarding Heroin incidences. It may be noted that heroin and opioid painkillers are extremely similar in terms of their chemical structure, mechanism of action and range of effects. Accordingly, for the purpose of this study, we use Heroin and other Opioid drug related data, in a similar fashion.

3.1.7 DS7: This dataset contains information regarding the median income, median age and educational distribution of various neighborhoods of Cincinnati. Information about the median income, median age and educational distribution were mined from three separate websites [7, 8].

3.2 Data Processing and Munging

We process and munge data from our collected datasets DS1-DS7, to create "secondary" datasets DS8-DS15 for the purpose of identification of at-risk groups, in the state of Connecticut. In the following, we describe our munging process:

3.2.1 DS8: It is a subset of DS1, restricted by the year 2016, to make it consistent with DS2.

3.2.2 DS9: It comprises of records of DS8, sorted by the counties, and filtered by gender, race and age.

3.2.3 DS10: It is created from DS9 and consists of 8 rows (corresponding to eight counties) and 13 columns (corresponding to number of incidences in each county; gender - Male, Female, Other; race - White, African American, Hispanic/Latino, Asian, other; age - below 20, 20-39, 40-59, above 60).

3.2.4 DS11: It is created from DS10 by considering all possible combinations of County, Gender, Race and Age. Since we have 8 different counties, 3 different genders and 5 different races and 4 different age levels, we will have 480 rows (8 * 3 * 5 * 4) and 5 columns (corresponding to a county, gender, race, age and the number of incidences for that specific county, gender, race and age).

3.2.5 DS12: It is created by sorting DS11 in descending order of the number of incidences.

3.2.6 DS13: This dataset is created by processing information available in DS4 and DS5. From DS4, we create a temporary dataset DS4A that contains information regarding the total number of prescribers and prescriptions written in each of the 50 states. DS4A was *joined* with DS5, to create DS13, that contains information regarding the total number of prescribers, prescriptions and Opioid related deaths in each of the 50 states.

3.2.7 DS14: This dataset was created by processing information available in DS6, and it contains information related to the number of Opioid related incidences in each of the 50 neighborhoods of Cincinnati.

3.2.8 DS15: This dataset was created by processing information available in datasets DS7 and DS14 and it contains information related to the median income, median age, *median education* and the number of Opioid related incidences in each of the 50 neighborhoods of Cincinnati. It may be noted that DS7 provides information related to the distribution of educational level of each of the neighborhoods. We define median education level of a neighborhood as the number of years, 50% of the residents of the neighborhood spend in school. In [8], the educational level is divided into 10 different categories from c_1, ..., c_{10} where c_1 corresponds to *None* and c_{10} corresponds to *Doctorate*. The categories c_1, ..., c_{10} correspond to n_1, ..., n_{10} years of education, with *None* implying 0 years of education and *Doctorate* implying 22 years of education. The precise definition of median education level of a neighborhood is as follows. The median educational level of a neighborhood is n_k years, if k is the smallest integer, such that $\sum_{i=1}^{k} x_i \geq 50$, where x_1, ..., x_{10} represents the percentage of neighborhood population that has educational levels corresponding to c_1, ..., c_{10}.

4 PROPOSED WORK

In order to identify at-risk groups with a high level of accuracy, one obviously needs to have access to relevant data. Data pertaining to Opioid related prescriptions, incidences, such as, calls to Emergency Medical Services (EMS) and deaths, are owned by multiple stakeholders, such as the insurance companies, hospitals, EMS providers and drug stores. Public health organizations at the federal, state and local level often collect such data, anonymize and aggregate them and make it available on the web. For our analysis, we have used such data made available by Centers of Medical and Medicaid Services [3], Health and Human Services department of Connecticut [1] and EMS responses for the city of Cincinnati [6]. However, such data often do not contain information at the individual level, e.g., it does not provide medical history of an individual as to how many or how often the individual was taking Opioid drugs. Moreover, it does not contain socio-economic and educational background of the individual. Data related to the medical history of an individual is available to the insurance companies, hospitals and drug stores. We did not have access to such data. However, we are making an effort to collect such data from these sources.

Once we acquire such data, we plan to develop a mathematical model to estimate the association between Opioid abuse and medical history and demographic characteristics of individuals. Given the demographic information and medical history of an individual, the response variable of the model, will assign the individual to a risk group. We also plan to develop unsupervised learning paradigms, as the number of risk groups may vary over time. We are currently developing mixture models, using the Expectation-Maximization algorithm.

5 PRELIMINARY WORK RESULTS

In this section, we present preliminary results for identification of at-risk groups, as well as answers to Q1-Q3, presented earlier. In the following, we briefly discuss these results

5.1 At-Risk Group Identification

As noted earlier, ideally, we would have liked to identify at-risk groups at the following level of granularity and able to make statements of the form: "White Male *residents of* Fairfield County *in the* age group of 20-39*, within* income bracket $30k-$60k*, and* educational level high school graduate*"*, are the highest risk group in Connecticut. However, the datasets available to us currently does not provide any information related to income and education level of the individual involved in the Opioid related incidence. Accordingly, we are unable to identify at-risk groups at a level of

granularity involving six factors (Race, Gender, County, Age, Income and Education). Instead, we identified at-risk groups involving four factors (Race, Gender, County, Age).

It may be recalled from Section. 3.2 that DS12 was created by sorting DS11 in descending order of the number of incidences, where DS11 contained all possible combinations of County, Gender, Race and Age. Thus, DS12 contains 480 rows (corresponding to 8 different counties, 3 different genders and 5 different races and 4 different age levels) and 5 columns (corresponding to county, gender, race, age and the number of incidences for that specific county, gender, race and age). Each of the 480 rows correspond to a *risk group* identified by county, gender, race and age. We define a risk group to be the *highest risk group* if the number of Opioid related incidences for this group is the highest among all risk groups. Based on our analysis, we present the five highest risk groups, in Connecticut, in Table. 1.

County	Race	Gender	Age Group	No. of Incidences
Hartford	White	Male	20-39	66
New Haven	White	Male	40-59	64
Hartford	White	Male	40-59	61
New Haven	White	Male	20-39	53
Fairfield	White	Male	20-39	44

Table 1: Five Highest Risk Groups in Connecticut

From Table. 1, we find that White Males as a demographic group, is the highest risk group in Connecticut. Moreover, two counties - Hartford and New Haven, are among the worst affected, by the Opioid epidemic. From a different dataset DS3, we can identify income and educational level characteristics of the highest risk groups (White, Male, Hartford/New Haven County). The income level for White males in Hartford, Connecticut was $63,200 in 2016 and 91.7% of them were at least High School graduates. The corresponding numbers for New Haven, Connecticut were $48,985 in 2016 and 91.3% respectively.

5.2 Results of Data Analysis for Q1-Q3

We used DS4, to answer the first two questions. Question3 was answered using DS15. We briefly summarize our findings below,

5.2.1 Data Analysis for Q1. We used partial correlation to analyze the relationship between the prescribers/number of prescriptions and the number of Opioid related deaths. From Table. 2, we can infer that there is a moderate positive correlation between the number of prescribers and prescriptions with Opioid related deaths.

	Number of Prescribers	Number of Prescriptions
Opiate Deaths (Partial Correlation)	0.4664	0.3619

Table 2: Partial Correlation Coefficients between Opiate Deaths and Prescribers and Prescriptions

5.2.2 Data Analysis for Q2. We used several machine learning algorithms to predict prescribers who are likely to prescribe high number of Opioid prescriptions, by analyzing the trend of prescribing *non Opioid* drugs. IBM ran some initial machine learning algorithms and attained accuracies ranging from 60% to 84%. Using XGBoost and CatBoost, we had accuracy scores of 81.8% and 84.7%. The CatBoost model provided a feature importance array, which identified "specialty" as the most important feature. The prescribers with specialty "Addictive Medicine", prescribed the highest average of annual Opioid drugs. We also examined the average annual Opioid prescription rates by state and observed a trend of higher prescription rates in the southern states. Furthermore, we implemented a Mulit-Layer Perceptron and a Random Forest on the dataset. We did not consider the specialty of the prescribers in these two models. We achieved a training accuracy of 95.6% and a testing accuracy of 89.7%, using the MLP, and a testing accuracy of 89% using the Random Forest model.

5.2.3 Data Analyses for Q3. We used the Cincinnati Heroin Overdose dataset, along with the Cincinnati neighborhood dataset to identify the relationship (using partial correlation) between the income levels, age and educational levels of a neighborhood, and the number of Opioid related incidences. The results are tabulated in Table. 3. From the table, we can infer that the Opioid addiction affects the entire spectrum of income levels and age, and is not restricted to a particular level. In addition, we found that, with an increase in the educational level of a neighborhood, there is a decrease in the Opioid related deaths, albeit slightly.

	Median Income	Median Age	Median Education
Opiate Deaths (Partial Correlation)	-0.0576	-0.0789	-0.1516

Table 3: Partial Correlation Coefficients between Opiate Deaths and Median Income/Age/Education

REFERENCES

[1] City of Connecticut, "Accidental Drug Related Deaths", https://data.ct.gov/Health-and-Human-Services/Accidental-Drug-Related-Deaths-2012-2017/ecj5-r2i9
[2] USDA, Economic Research Service, "https://www.ers.usda.gov/data-products/county-level-data-sets/", 2016
[3] Centers for Medicare and Medicaid Services, "https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html.
[4] Kaggle Dataset: "U.S. Opiate Prescriptions/Overdoses", https://www.kaggle.com/apryor6/us-opiate-prescriptions.
[5] IBM Opioid Github: "https://github.com/IBM/predict-opioid-prescribers", 2017.
[6] City of Cincinnati, "Heroin Overdoses", https://insights.cincinnati-oh.gov/stories/s/Heroin/dm3s-ep3u/.
[7] City-Data, "http://www.city-data.com/city/Cincinnati-Ohio.html".
[8] Statistical Atlas, "https://statisticalatlas.com/place/Ohio/Cincinnati/Overview".
[9] J. B. Rice, A. G. White, H. G. Birnbaum, M. Schiller, D. A. Brown, and C. L. Roland. "A model to identify patients at risk for prescription opioid abuse, dependence, and misuse." Pain Medicine 13, no. 9 (2012): 1162-1173.
[10] R. Chou, J.A. Turner, E. B. Devine, R. N. Hansen, S. D. Sullivan, I. Blazina, T. Dana, C. Bougatsos, and R. A. Deyo. "The effectiveness and risks of long-term opioid therapy for chronic pain", Annals of internal medicine 162, no. 4 (2015): 276-286.
[11] D. B. Neill, W. Herlands. "Machine Learning for Drug Overdose Surveillance." Journal of Technology in Human Services (2018): 1-7.
[12] T.K. Mackey, J. Kalyanam, T. Katsuki, G. Lanckriet, "Twitter-Based Detection of Illegal Online Sale of Prescription Opioid", American J. of Public Health, 2017.
[13] M. Chary, N. Genes, C. Giraud-Carrier, C. Hanson, L.S. Nelson, A.F. Manini, "Epidemiology from Tweets: Estimating Misuse of Prescription Opioids in the USA from Social Media", Journal of Medical Toxicology, 13, 278-286, 2017.
[14] D. Wei, "Combating the Opioid Epidemic with Machine Learning", https://www.ibm.com/blogs/research/2017/08/combating-the-opioid-epidemic-with-machine-learning/, 2017.
[15] United States Census Bureau Fact Finder, "https://factfinder.census.gov", 2016.

Epidemiological Data and Model Requirements to Support Policy

Marc Baguelin
Public Health England
London, UK
marc.baguelin@phe.gov.uk

Elizabeth Buckingham-Jeffery
School of Mathematics
Manchester, UK
e.buckingham-jeffery@manchester.ac.uk

Ian Hall
School of Mathematics
Manchester, UK
ian.hall@manchester.ac.uk

Thomas House
School of Mathematics
Manchester, UK
thomas.house@manchester.ac.uk

Timothy Kinyanjui
School of Mathematics
Manchester, UK
timothymuiruri.kinyanjui@manchester.ac.uk

Lorenzo Pellis
School of Mathematics
Manchester, UK
lorenzo.pellis@manchester.ac.uk

ABSTRACT

Often, the task of epidemiological modelling is seen as one of improving biological and social realism, with increasing data availability enabling increased realism. Here, we consider ways in which models designed to support policy have different data and algorithmic requirements from those aiming at realism or insight, via a series of case studies. In particular, calculation and communication of uncertainty is often more important than refinement of model structure without confirmation of validity.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; • **Computing methodologies** → *Artificial intelligence*; *Modeling and simulation*;

KEYWORDS

Epidemic; Disease Dynamics; Uncertainty Quantification; Inference

ACM Reference Format:
Marc Baguelin, Elizabeth Buckingham-Jeffery, Ian Hall, Thomas House, Timothy Kinyanjui, and Lorenzo Pellis. 2018. Epidemiological Data and Model Requirements to Support Policy. In *Proceedings of epiDAMIK, ACM SIGKDD (epiDAMIK).* ACM, New York, NY, USA, Article tbc, 5 pages. https://doi.org/tbc

1 OVERVIEW

Epidemiology is a data-driven science; however, the sources of data involved are typically observational, since controlled experiments are seldom practical. This means mathematical modelling has a key role to play in the field, both in terms of inference and of prediction [17].

Inference is challenging because data is usually scarse and only indirectly informing our knowledge, as most events involved in the transmission process (e.g. infection, or beginning/end of viral shedding, which does not necessarily correspond to onset/end of symptoms) are rarely observed. Models offer an opportunity to codify our biological understanding (or belief) about the infection mechanism, but their integration with data requires sophisticated statistical techniques, particularly those that successfully cope with missing data.

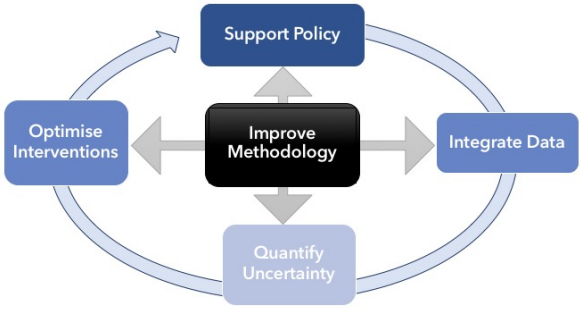


Figure 1: Representation of an idealised modelling support cycle for policy.

Models are also key for predictions, as they allow for the representation of phenomena such as potential interventions that are not directly observable. Unfortunately, validation of model predictions is problematic, as no epidemic is ever identical to any other and testing alternative control policies is anyway constrained by ethical or political considerations.

When modelling to inform policy, particularly when a decision must be made under time pressure, there are particular practical and theoretical challenges. These are often markedly different from those that arise in curiosity-driven science. Figure 1 shows an overall picture of modelling for policy support. A successful policy-driven model will integrate available data, and propagate forward the most significant uncertainties to allow interventions to be optimised, all the while improving methodology. The rest of this position paper is structured as a set of case studies that illustrate this point.

epiDAMIK, August 2018, London, UK

ACM ISBN tbc.
https://doi.org/tbc

2 TYPES OF MODELS USED

Broadly speaking, there are two classes of models (although these can be related to each other and embedded in more general frameworks) that we now present simple examples of for clarity.

The first of these is the Poisson process for a non-communicable disease. Here we imagine that we observe a number of cases of disease y in a population of size N over a time period of length t. In the simplest model, we assume that $N \gg y$ and that cases arise independently at a rate λ, leading to a Poisson likelihood of

$$\Pr(y) = \frac{(\lambda t)^y \mathrm{e}^{-\lambda t}}{y!} \,. \tag{1}$$

This Poisson model, if correct, would allow us to estimate λ from data (for example if 365 cases are observed in a year then we would estimate the rate to be $\hat{\lambda} = 1$ days^{-1}) and on that basis could use (1) to predict the probability that we observe significantly more cases than expected, leading to a lack of healthcare capacity, on a given week.

The second is the SIR model, in which individuals are split into compartments according to whether they are Susceptible to the disease, Infectious, or Removed. This model is often represented through the non-linear system of differential equations:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\beta}{N} SI \,; \qquad \frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\beta}{N} SI - \gamma I \,. \tag{2}$$

If we know the values $S(t = 0)$ and $I(t = 0)$, as well as the mean duration of infectiousness $1/\gamma$ and the basic reproductive ratio $R_0 = \beta/\gamma$ then this model (and related approaches [23]) can be used to calculate quantities such as the value of t at which $I(t)$ is maximised – i.e. the peak time and height – as well as the total number of individuals infected during the epidemic, $N - S(\infty)$.

Elaborating on the back-bone of the SIR model (2), a range of more sophisticated model structures has been proposed to relax unrealistic assumptions and capture various aspects of human social patterns, including: multitype models that distinguish between classes of individuals (e.g. age or risk behaviour); metapopulation models (e.g. cities connected by flights); households models, capturing social grouping imposed by household, school and workplace structures; network models; spatial models, e.g. incorporating transmission reduction with distance or movements dependent on population density; and complex individual-based stochastic simulations, which are flexible but involve many parameters. For reviews, see Rock et al. [37] and Keeling and Rohani [23].

3 REAL-TIME DECISION MAKING DURING PANDEMICS

Infectious disease modelling is increasingly used to support decision making in real-time in order to choose best control strategies. In 2001, for example, modelling was used to estimate the impact of potential control strategies during the outbreak of foot-and-mouth disease in the UK [14, 24]. During the last influenza pandemic in 2009, a cost-effectiveness analysis involving a transmission model fitted to incoming data informed the UK government on the likely impact of alternative vaccination strategies [2]. More recently, during the Ebola outbreak in West Africa, models were used to forecast the likely course of the outbreak [41], evaluate the benefits and risks of introducing Ebola community centres [26], estimate the impact of new beds [6] and evaluate clinical trials for experimental treatments [11].

In these situations, trying to develop increasingly more complex models in order to integrate additional parameters relevant to decision makers is often not possible using traditional methods. The reason is that traditional methods of evidence synthesis are based on computationally intensive algorithms poorly adapted to quick responsive real-time inference. The quality of real-time data involved integrating additional modelling layers to reflect potential censoring and sources of uncertainties. Finally, these models need to incorporate simple summary outputs which can be handled during decision making.

A real-time modelling toolbox needs to be developed to provide a set of modular methodologies which can be picked up to build a model flexible enough to integrate relevant complexities while still being fitted in a short amount of time. Such tools could involve heavily parallelised methods that exploit multi-core computer architecture such as particle filters, variational Bayesian approaches, or approximation [5]. It is, however, possible that any computational gain could be offset by approximation error. It would thus also be necessary to develop, prior to these crises, studies quantifying the level of bias introduced by using 'fast' methods.

4 PREDICTING THE PEAK DEMAND FOR HEALTHCARE

Prediction of height and timing of peak incidence is crucial to estimate the stress on the health care system, and hence to inform decision-makers on how to allocate resources to manage the outbreak most cost-effectively. Hospitals running out of beds during more severe influenza seasons are not uncommon [7, 8], with patient care severely delayed and other hospital services postponed to after the winter crisis. Conversely, enough resources in terms of bed capacity in treatment centres has been shown to have contributed to controlling the 2014 Ebola outbreak [27].

Given the abundance of studies discussing, comparing and testing methods for prediction of influenza epidemics [32, 34, 40, 44], we focus here on the case of influenza, although comments naturally extend to other infections. Among the many epidemiologically relevant epidemic characteristics [40], peak timing and height are among the most widely considered [32, 33, 38].

Stochastic simulations of simple epidemic models based on the 'mass-action' mixing assumption highlight how peak epidemic timing can vary widely as a consequence of the random delays in the early epidemic phase. However, once the epidemic takes off and the number of cases becomes large enough to motivate a deterministic approximation, the explosive nature of exponential growth is such that uncertainties regarding the size of the population under consideration or the fraction effectively susceptible to infection has only marginal impact on how quickly the peak is reached. Mathematically speaking, the stochastic counterpart of model (2), if not for the initial and final epidemic phases dominated by random events, consists of a 'deterministic' central phase the duration of which is $O(1)$ (i.e. independent of the population size N) [1]. It is unsurprising, therefore, that numerous retrospective studies of influenza concluded peak timing prediction can be already accurate weeks in advance (cited numbers generally range form 4 to 7 [32, 33, 38]), at

least in those years characterised by a single epidemic wave and provided enough data is promptly available. Prospective, real-time forecasts [39, 42] also proved reasonably accurate even up to 9 weeks in advance.

The delays between infections and reliable data on confirmed cases becoming available (1-2 weeks [34] or even longer [9]) may threaten the usefulness of fast prediction methods. Alternatives using more promptly available data, such as influenza-like illness, acute respiratory infection data or absenteeism, or data surrogates such as Google 'Flu Trends and search engine queries, unavoidably inherit the inaccuracies of such data sources. For example, the wide media coverage during the 2012-2013 epidemic in the USA has been claimed as a potential cause for the time discrepancy between Google search activity and the peak in real infections [39]. However, forecast methods based on combinations of social (e.g. Internet searches) and physical indicators (e.g. absolute humidity) appear to outperform those based on single indicators alone, and can be further strengthened by accounting for uncertainty due to offical estimates undergoing revision after publication [9].

If peak timing predictions appear broadly successful, the prediction of peak height is much more challenging [32, 34]. A potential explanation is that, even when mass-action mixing is a reasonable assumption, the height of the peak is highly dependent on the exact speed of epidemc growth [32] and other factors that are hard to estimate, such as: the distribution of (often partial) susceptibility in the population, the rate of under-reporting, and the effect of control policies (e.g. vaccination) and behavioural change (e.g. self-quarantining, reduced mixing, but also "flu parties" [29]).

When the mass-action assumption is not justifiable, models with a more complex structure might need to be employed, complicating the matter further. In a metapopulation framework, the height and timing of peak incidence in the full population results from the superposition of the subpopulation dynamics. As such, they depend on subpopulation sizes, but also, crucially, on the times at which the epidemic jumps between subpopulations, which are often 'rare' events that are hard to predict accurately. However, because as discussed above peak timing can be predicted in each subpopulation provided local data is promptly available, the work of Shaman et al. [39], which ignores a metapopulation structure and treats different cities as independent of each other, could still forecast peak timing for many of the 108 cities considered.

Forecasting methods typically involve complex simulation models and relatively simple statistical fitting procedures [33] or relatively simple models whose potential misspecification is compensated for by sophisticated statistical approaches, such as particle filters [32] and data assimilation ensemble approaches originally adopted in numerical weather prediction [38, 39]. No single method appears uniformly better than others and even after the peak has passed there are instances when peak timing forecasts are inaccurate [44]. In particular, all the methodologies tested in [44] struggled in performing prediction for those years characterised by two (or even three) separate peaks, as they were the result of separate outbreaks of different strains of influenza that the models were not designed to capture. Integrating the sophisticated statistical techniques with more realistic models of seasonal influenza is a challenge for the future, but the lack of accurate and promptly available data remains one of the main limitations in epidemiology, especially compared to weather and climate predictions [28].

5 SYNDROMIC SURVEILLANCE FOR SITUATIONAL AWARENESS

Syndromic surveillance is the monitoring of the number of cases of illness in a population with a specified syndrome [36]. Priorities for the algorithms used to monitor real-time syndromic signals are robustness, speed, and the ability to model signals at different scales (some signals have many days of zero cases and some have many cases every day).

Methods used in practice to monitor signals such as vomiting or influenza-like illness at different points in the healthcare system include the 'early aberration reporting system' developed by the CDC [21], the 'moving epidemic method' [43], and the 'rising activity, multi-level mixed effects, indicator emphasis' (RAMMIE) method developed by Public Health England [31]. These are statistical methods not based on dynamical systems of transmissionthat are not transmission dynamic. RAMMIE, for example, is based on a Poisson model (1) with an additional multi-level structure to exploit signals from hierachical geographies (national, regional, and local levels) [31].

Such approaches, applied to daily data, can improve situational awareness during mass gatherings [18], provide support during environmental problems [13], and detect and follow trends in larger seasonal influenza outbreaks [20]. However, smaller outbreaks, for example small gastrointestinal outbreaks [10] or anthrax releases [30], are unlikely to be detected; the possibility of greater sensitivity therefore remains open.

6 RESPONDING TO DISEASE OUTBREAKS WITH ENVIRONMENTAL SOURCES

Diseases which arise from environmental sources rather than by person to person spread – for example anthrax and Legionnaires disease as opposed to measles or influenza – do not have the nonlinearity inherent in SIR-type models. This makes them mathematically simpler but still have their own modelling challenges [12]. These outbreaks will involve fewer people than expected from a pandemic influenza wave but may be deliberate or arise during high profile events such as the Olympic Games meaning the timescales for decisions are shorter and uncertainty in data greater.

As was shown in [30] traditional syndromic surveillance schemes are unlikely to detect the atypical emerging disease outbreaks. Instead such infections are likely to be detected within hospital settings. This means the key data required to run models must be collected on the fly from cases. Given the speed of data collection there is likely to be uncertainty in quality of data. Efficient data transfer is essential (using electronic data capture software and transfer formats [15]) and wider adoption of such technology by responders is key, where practical.

As shown in [16], methods may be adapted on the fly to support outbreaks. Work between outbreaks can extend methods more robustly (by allowing for additional delays or multiple sources) but every outbreak is unique and so the perfect model design and data demands are difficult to write down *a priori*.

A key uncertainty is the location of people at the time of infection and the number of people present that did not get infected. Traditional epidemiological methods of interviewing observed cases mean that one can reasonably expect home and work location of cases. This may be matched to home and work locations of the general population from surveys or censuses but if infection arises during travel or leisure activity the infection location may be missed. Even with more detailed travel history from cases this denominator population is hard to define. In [19] the authors have looked at transient movements using novel data set including detailed travel histories for defining denominators but such commercial datasets are expensive to maintain and often appear as 'black boxes' to eventual model users. To be viable for translation to public health organisations the tools must be affordable and transparent. Emerging data sources such as mobile phone location data may be a benefit to public health authorities, to define the at risk population, but access to such data is hard to ensure (particularly ahead of time so methods can be developed) and accuracy is hard to quantify.

Whilst inferential methods for reverse epidemiology (finding exposure locations given cases) exist [12], further methodological development is required, for example fusing intervention models [35] with inference tools to ensure realistic modelling of interventions. Futhermore, given that evidence for human dose response arises from animal studies and for relatively high doses, the infection of humans receiving low doses of biological agent is uncertain.

Given uncertainty in the models, and their justification, another challenge is around model selection (in absence of clear nesting of models) and presentation of key assumptions in an intuitive way to a lay audience. Models may have challenges in explaining concepts to such audiences but are critical in such situations to support outbreak control team work.

7 OUTBREAK CONTROL IN CARE HOMES

Care homes are an integral setting for disease transmission and control, especially within the context of an ageing population. A recent modelling study of scabies in care homes [25] suggested that early detection of an infectious index case is critical in establishing who and when to treat. Uncertainty in determining the index case leads to possibly a greater number of residents being infected with the potential of infecting staff who maintain frequent links between the care home and the general population.

Traditionally, the study of transmission of an infectious agent within such a setting has been studied using either an agent-based simulation model or a mean-field model [23]. For the sake of argument and without loss of generality, a care home can be regarded as a large household. These models capture the complete range of stochastic behaviours using a large set of ordinary differential equations which can be expressed succinctly as

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = Q\mathbf{p}\,, \qquad \mathbf{p}(0) = \mathbf{p}_0\,, \tag{3}$$

where $Q \in \mathbb{R}^{n\times n}$ is the household transition matrix and $\mathbf{p}$ is the probability that a household is in a certain infection configuration. For example, if we consider the SIR model as in (2) then we would have $\mathbf{p}_{s,i,r}(t)$ being the probability that a household has s susceptibles, i infectious and r removed. This is made possible because we count the number of events of each type that can occur rather than keep track of the population numbers. For example, in the SIR model only two events can occur, namely infection $(S, I, R) \to (S-1, I+1, R)$ and recovery $(S, I, R) \to (S, I-1, R+1)$. For a detailed overview of this type of models, see [3, 4]. The solution for (3) involves the exponential of a matrix:

$$\mathbf{p}(t) = \exp(tQ)\mathbf{p}_0\,. \tag{4}$$

To fully achieve the modelling ideal of sufficiently accounting for uncertainty, in both structure and parameter values, and being able to propagate it within a modelling framework, computationally efficient algorithms for solving the master equation are needed. Jenkinson and Goutsias [22] have considered an implicit Euler implementation to approximate the solution of (4). More recently, Kinyanjui et al. [25] have demonstrated that there exist computational advantages in solving the matrix exponential using expansion-based methods for the time-homogeneous case, i.e. when Q is time-invariant. The computational advantages gained allowed for a full Bayesian quantification of scabies transmission and control within households fully accounting for uncertainty [25]. However, there is still an open research question as to what happens when we have interactions between households making Q temporally heterogeneous.

8 CONCLUSIONS

In summary, we have outlined a series of policy-driven modelling contexts where data, time pressures and uncertainty limit the detail that can be incorporated into models, highlighting where further – particularly methodological – work is needed.

REFERENCES

[1] Håkan Andersson and Tom Britton. 2000. *Stochastic Epidemic Models and Their Statistical Analysis*. Springer Lectures Notes in Statistics, Vol. 151. Springer, Berlin.

[2] Marc Baguelin, Albert Jan Van Hoek, Mark Jit, Stefan Flasche, Peter J. White, and W. John Edmunds. 2010. Vaccination against pandemic influenza A/H1N1v in England: A real-time economic evaluation. *Vaccine* 28, 12 (mar 2010), 2370–2384. https://doi.org/10.1016/j.vaccine.2010.01.002

[3] Frank G Ball and Owen D. Lyne. 2001. Stochastic Multitype SIR Epidemics among a Population Partitioned into Households. *Advances in Applied Probability* 33, 1 (2001), 99–123. http://projecteuclid.org/euclid.aap/999187899

[4] Andrew J. Black and Joshua V. Ross. 2015. Computation of epidemic final size distributions. *Journal of Theoretical Biology* 367 (2015), 159–165. https://doi.org/10.1016/j.jtbi.2014.11.029 arXiv:arXiv:1407.3887v2

[5] Elizabeth Buckingham-Jeffery, Valerie Isham, and Thomas House. 2018. Gaussian process approximations for fast inference from infectious disease data. *Mathematical Biosciences* 301 (2018), 111–120.

[6] Anton Camacho, Adam Kucharski, Yvonne Aki-Sawyerr, Mark A White, Stefan Flasche, Marc Baguelin, Timothy Pollington, Julia R Carney, Rebecca Glover, Elizabeth Smout, et al. 2015. Temporal changes in Ebola transmission in Sierra Leone and implications for control requirements: a real-time modelling study. *PLoS currents* 7 (2015).

[7] Denis Campbell, Pamela Duncan, and Sarah Marsh. 2018. NHS patients dying in hospital corridors, A&E doctors tell Theresa May. https://www.theguardian.com/society/2018/jan/11/nhs-patients-dying-in-hospital-corridors-doctors-tell-theresa-may. First published: 2018-01-11.

[8] Denis Campbell and Sarah Marsh. 2017. NHS crisis: 20 hospitals declare black alert as patient safety no longer assured. https://www.theguardian.com/society/2017/jan/11/nhs-crisis-20-hospitals-declare-black-alert-as-patient-safety-no-longer-assured. First published: 2017-01-11.

[9] Prithwish Chakraborty, Pejman Khadivi, Bryan Lewis, Aravindan Mahendiran, Jiangzhuo Chen, Patrick Butler, Elaine O Nsoesie, Sumiko R Mekaru, John S

Brownstein, Madhav V Marathe, et al. 2014. Forecasting a moving target: Ensemble models for ILI case count predictions. In *Proceedings of the 2014 SIAM international conference on data mining*. SIAM, 262–270.

[10] Felipe J. Colón-González, Iain R. Lake, Roger A. Morbey, Alex J. Elliot, Richard Pebody, and Gillian E. Smith. 2018. A methodological framework for the evaluation of syndromic surveillance systems: a case study of England. *BMC Public Health* 18, 1 (2018), 544. https://doi.org/10.1186/s12889-018-5422-9

[11] Ben S. Cooper, Maciej F. Boni, Wirichada Pan-ngum, Nicholas P. J. Day, Peter W. Horby, Piero Olliaro, Trudie Lang, Nicholas J. White, Lisa J. White, and John Whitehead. 2015. Evaluating Clinical Trial Designs for Investigational Treatments of Ebola Virus Disease. *PLOS Medicine* 12, 4 (04 2015), 1–14. https://doi.org/10.1371/journal.pmed.1001815

[12] Joseph R. Egan and Ian M. Hall. 2015. A review of back-calculation techniques and their potential to inform mitigation strategies with application to non-transmissible acute infectious diseases. *Journal of The Royal Society Interface* 12, 106 (2015).

[13] Gillian E.Smith, Zharain Bawa, Yolande Macklin, Roger Morbey, Alec Dobney, Sotiris Vardoulakis, and Alex J.Elliot. 2015. Using real-time syndromic surveillance systems to help explore the acute impact of the air pollution incident of March/April 2014 in England. *Environmental Research* 136 (2015), 500–504.

[14] Neil M. Ferguson, Christl A. Donnelly, and Roy M. Anderson. 2001. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature* 413, 6855 (oct 2001), 542–548. https://doi.org/10.1038/35097116

[15] Thomas J.R. Finnie, Andy South, Ana Bento, Ellie Sherrard-Smith, and Thibaut Jombart. 2016. EpiJSON: A unified data-format for epidemiology. *Epidemics* 15 (2016), 20–26.

[16] Maya Gobin, Jeremy Hawker, Paul Cleary, Thomas Inns, Daniel Gardiner, Amy Mikhail, Jacquelyn McCormick, Richard Elson, Derren Ready, Tim Dallman, Iain Roddick, Ian Hall, Caroline Willis, Paul Crook, Gauri Godbole, Drazenka Tubin-Delic, and Isabel Oliver. 2018. National outbreak of Shiga toxin-producing Escherichia coli O157:H7 linked to mixed salad leaves, United Kingdom, 2016. *Eurosurveillance* 23, 18 (2018).

[17] Nicholas C Grassly and Christophe Fraser. 2008. Mathematical models of infectious disease transmission. *Nature Reviews Microbiology* 6, 6 (2008), 477.

[18] Adi V. Gundlapalli, Jonathan Olson, Sean P. Smith, Michael Baza, Robert R. Hausam, Louise J. Eutropius, Stanley L. Pestotnik, Karen Duncan, Nancy Staggers, Pierre Pincetl, and Matthew H. Samore. 2007. Hospital electronic medical record-based public health surveillance system deployed during the 2002 Winter Olympic Games. *American Journal of Infection Control* 35 (2007), 163–171. Issue 3.

[19] Penelope A. Hancock, Yasmin Rehman, Ian M. Hall, Obaghe Edeghere, Leon Danon, Thomas A. House, and Matthew J. Keeling. 2014. Strategies for Controlling Non-Transmissible Infection Outbreaks Using a Large Human Movement Data Set. *PLOS Computational Biology* 10, 9 (2014), e1003809. https://doi.org/10.1371/journal.pcbi.1003809

[20] S. E. Harcourt, G. E. Smith, A. J. Elliot, and R. Pebody. 2012. Use of a large general practice syndromic surveillance system to monitor the progress of the influenza A(H1N1) pandemic 2009 in the UK. *Epidemiology and Infection* 140 (2012), 100–105.

[21] Lori Hutwagner, William Thompson, G Matthew Seeman, and Tracee Treadwell. 2003. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *Journal of urban health : bulletin of the New York Academy of Medicine* 80, 2 Suppl 1 (2003), i89–i96. https://doi.org/10.1007/PL00022319

[22] Garrett Jenkinson and John Goutsias. 2012. Numerical integration of the master equation in some models of stochastic epidemiology. *PloS one* 7, 5 (jan 2012), e36160. https://doi.org/10.1371/journal.pone.0036160

[23] Matt J Keeling and Pejman Rohani. 2007. *Modeling infectious diseases in humans and animals*. Princeton University Press.

[24] Matt J Keeling, Mark EJ Woolhouse, Darren J Shaw, Louise Matthews, Margo Chase-Topping, Dan T Haydon, Stephen J Cornell, Jens Kappey, John Wilesmith, and Bryan T Grenfell. 2001. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* 294, 5543 (2001), 813–817.

[25] Timothy Kinyanjui, Jo Middleton, Stefan Güttel, Jackie Cassell, Joshua Ross, and Thomas House. 2018. Scabies in residential care homes: Modelling, inference and interventions for well-connected population sub-units. *PLOS Computational Biology* 14, 3 (2018), e1006046.

[26] Adam A.J. Kucharski, Anton Camacho, Francesco Checchi, Ron Waldman, Rebecca R.F. Grais, Jean-Clement J.-C. Cabrol, Sylvie Briand, Marc Baguelin, Stefan Flasche, Sebastian Funk, William John Edmunds, and W. John Edmunds. 2015. Evaluation of the Benefits and Risks of Introducing Ebola Community Care Centers, Sierra Leone. *Emerging infectious diseases* 21, 3 (2015), 393–399. https://doi.org/10.3201/eid2103.141892

[27] Adam J Kucharski, Anton Camacho, Stefan Flasche, Rebecca E Glover, W John Edmunds, and Sebastian Funk. 2015. Measuring the impact of Ebola control measures in Sierra Leone. *Proceedings of the National Academy of Sciences* 112, 46 (2015), 14366–14371.

[28] Tom Lindström, Michael Tildesley, and Colleen Webb. 2015. A bayesian ensemble approach for epidemiological projections. *PLoS computational biology* 11, 4 (2015), e1004187.

[29] Donald G. McNeil Jr. 2009. Debating the Wisdom of 'Swine Flu Parties'. https://www.nytimes.com/2009/05/07/world/americas/07party.html. First published: 2009-05-06.

[30] RA Morbey, AJ Elliot, A Charlett, S Ibbotson, NQ Verlander, S Leach, I Hall, I Barrass, M Catchpole, B McCloskey, et al. 2014. Using public health scenarios to predict the utility of a national syndromic surveillance programme during the 2012 London Olympic and Paralympic Games. *Epidemiology & Infection* 142, 5 (2014), 984–993.

[31] Roger A. Morbey, Alex J. Elliot, Andre Charlett, Neville Q. Verlander, Nick Andrews, and Gillian E. Smith. 2015. The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis' (RAMMIE) method for syndromic surveillance in England. *Bioinformatics* 31, 22 (2015), 3660–3665. https://doi.org/10.1093/bioinformatics/btv418

[32] Robert Moss, Alexander Zarebski, Peter Dawson, and James M McCaw. 2016. Forecasting influenza outbreak dynamics in Melbourne from Internet search query surveillance data. *Influenza and other respiratory viruses* 10, 4 (2016), 314–323.

[33] Elaine Nsoesie, Madhav Mararthe, and John Brownstein. 2013. Forecasting peaks of seasonal influenza epidemics. *PLoS currents* 5 (2013).

[34] Elaine O Nsoesie, John S Brownstein, Naren Ramakrishnan, and Madhav V Marathe. 2014. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other respiratory viruses* 8, 3 (2014), 309–316.

[35] Gabriel Rainisch, Martin I Meltzer, Sean Shadomy, William A Bower, and Nathaniel Hupert. 2017. Modeling Tool for Decision Support during Early Days of an Anthrax Event. *Emerging Infectious Diseases* 23, 1 (2017), 46–55.

[36] Arthur Reingold. 2013. If syndromic surveillance is the answer, what is the question? *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 1, 2 (2013).

[37] Kat Rock, Sam Brand, Jo Moir, and Matt J Keeling. 2014. Dynamics of infectious diseases. *Reports on Progress in Physics* 77, 2 (2014), 026602.

[38] Jeffrey Shaman and Alicia Karspeck. 2012. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences* 109, 50 (2012), 20425–20430.

[39] Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. 2013. Real-time influenza forecasts during the 2012–2013 season. *Nature communications* 4 (2013), 2837.

[40] Farzaneh Sadat Tabataba, Prithwish Chakraborty, Naren Ramakrishnan, Srinivasan Venkatramanan, Jiangzhuo Chen, Bryan Lewis, and Madhav Marathe. 2017. A framework for evaluating epidemic forecasts. *BMC infectious diseases* 17, 1 (2017), 345.

[41] WHO Ebola Response Team. 2014. Ebola Virus Disease in West Africa – The First 9 Months of the Epidemic and Forward Projections. *New England Journal of Medicine* 371, 16 (oct 2014), 1481–1495. https://doi.org/10.1056/NEJMoa1411100

[42] Sherry Towers and Zhilan Feng. 2009. Pandemic H1N1 influenza: predicting the course of a pandemic and assessing the efficacy of the planned vaccination programme in the United States. *Eurosurveillance* 14, 41 (2009), 19358.

[43] Tomás Vega, Jose Eugenio Lozano, Tamara Meerhoff, René Snacken, Joshua Mott, Raul Ortiz de Lejarazu, and Baltazar Nunes. 2013. Influenza surveillance in Europe: Establishing epidemic thresholds by the Moving Epidemic Method. *Influenza and other Respiratory Viruses* 7, 4 (2013), 546–558. https://doi.org/10.1111/j.1750-2659.2012.00422.x

[44] Wan Yang, Alicia Karspeck, and Jeffrey Shaman. 2014. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS computational biology* 10, 4 (2014), e1003583.

Dynamics underlying global spread of emerging epidemics: An analytical framework

Lin Wang
WHO Collaborating Centre for Infectious Disease Epidemiology and Control, School of Public Health, Li Ka Shing Faculty of Medicine,
The University of Hong Kong
Hong Kong SAR, China
fdlwang@gmail.com

Joseph T Wu*
WHO Collaborating Centre for Infectious Disease Epidemiology and Control, School of Public Health, Li Ka Shing Faculty of Medicine,
The University of Hong Kong
Hong Kong SAR, China
joewu@hku.hk

ABSTRACT

Global spread of emerging epidemics (e.g. pandemic influenza, SARS, MERS-CoV, Ebola) is increasingly common, associated with the rapid pace of urbanization and global travel. Global metapopulation epidemic models built with worldwide air-transportation network (WAN) data have been one of the main tools for studying global spread of epidemics. However, it remains unclear how infectious disease epidemiology and the network properties of the WAN determine epidemic arrivals for different populations around the world. This work fills this knowledge gap by developing and validating an analytical framework on the basis of stochastic processes and network theory, which not only elucidates the dynamics underlying global spread of epidemics but also advances our capability in nowcasting and forecasting epidemics.

CCS CONCEPTS

• **Applied computing** → **Transportation**; **Forecasting**;
• **Mathematics of computing** → **Stochastic processes**;
• **Networks** → **Network dynamics**; **Network mobility**;
• **Computing methodologies** → **Network science**;

KEYWORDS

Emerging epidemics, Spatial epidemiology, Metapopulation epidemic models, Worldwide air-transportation network, Epidemic arrival time, Nonhomogeneous Poisson process

ACM Reference Format:
Lin Wang and Joseph T Wu. 2018. Dynamics underlying global spread of emerging epidemics: An analytical framework. In *Proceedings of The 2018 ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery (KDD2018).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Materials are available on request from the corresponding author: Joseph T Wu

KDD2018, August 2018, London, United Kingdom

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recent decades, global spread of emerging epidemics is increasingly common, as exemplified by the spread of SARS to nearly 30 countries in 2003, the spread of influenza A/H1N1 pandemic to more than 100 countries in 2009, the exportation of Ebola cases from the West Africa to the Nigeria, United States and United Kingdom in 2014, and recent geographical expansion of vector-borne diseases such as Dengue and Zika virus. Such frequent outbreaks of emerging epidemics are associated with the rapid pace of urbanization and global travel [5, 9, 14, 17]. In response to the serious situation, the World Health Organization (WHO) regularly updates the blueprint list of priority diseases to guide public health research and preparedness [28].

Since the 1980s, metapopulation epidemic models built with worldwide air-transportation network (WAN) data have been one of the main tools for studying global spread of emerging epidemics [12, 19, 24]. Despite their long history and widespread use, most studies in this field rely on computationally intensive simulations to predict or forecast the spatiotemporal transmission of epidemics [17, 19, 24]. However, one downside of such simulation-based methodology is that the computational process tends to be a black box – the underlying dynamics is hard to be elucidated from the basic principles in infectious disease epidemiology and network theory. In particular, an analytical understanding of the underlying dynamics has only been partially elucidated in recent years [4, 10, 23]. To fill this knowledge gap, we develop a novel analytical framework for characterizing how global spread of emerging epidemics depends on epidemiological parameters and the network properties of the WAN.

2 GLOBAL SPREAD SIMULATIONS: METAPOPULATION EPIDEMIC MODELS

2.1 Structure of the metapopulation epidemic models.

Metapopulation epidemic models are often described as a complex network of populations, in which each population denotes a city in the world and populations are interconnected through the mobility of individuals via the WAN [19, 24].

Since emerging infectious diseases generally evoke an epidemic with relatively fast timescales, we assume that in each population the epidemic peaks within 300 days after the establishment of the disease in that population [25]. It indicates that the change in demographics (e.g. births, aging) is negligible, such that each population has a constant population size. Denote population i as the epidemic origin with s_i initial infections seeded at time 0. For any given population j, the population size is denoted by N_j, with initial epidemic growth rate denoted by λ_j. For populations j and k that are directly connected, the per capita mobility rate from j to k is computed by $w_{jk} = F_{jk}/N_j$, in which F_{jk} is the daily number of passengers travelled by direct flights from j to k. Denote T_{ij}^n as the time at which population j receives its nth imported infection, such that T_{ij}^1 denotes the epidemic arrival time (EAT) for population j. **Table 1** summarizes the parameters.

2.1.1 Local epidemic dynamics within each population. The spread of epidemics within each population is modelled with frequency-dependent compartmental epidemic models [16], in which the transmission rate for infectious people to infect others can depend on multiple factors including the interpersonal contact rates, pathogenicity and environmental suitability [1, 6, 18, 31]. In the main text, we use the standard SIR model to describe the local epidemic dynamics within each population. Appendix A.1 extends to more general epidemic dynamics modelled by SE_mI_nR models.

Let $S_i(t)$, $I_i(t)$ and $R_i(t)$ be the number of susceptible, infectious and recovered people in a given population i at time t. Suppose $R_{0,i}$ is the basic reproductive number and $T_{g,i}$ is the mean generation time in population i. Let $\beta_i = R_{0,i}/T_{g,i}$ be the disease transmission rate and $\mu_i = 1/T_{g,i}$ be the recovery rate in population i. The SIR model is described by the following differential equations:

$$\frac{dS_i(t)}{dt} = -\beta_i \frac{S_i(t)}{N_i} I_i(t),$$
$$\frac{dI_i(t)}{dt} = \beta_i \frac{S_i(t)}{N_i} I_i(t) - \mu I_i(t),$$
$$\frac{dR_i(t)}{dt} = \mu I_i(t).$$

The doubling time $T_{d,i}$ for disease prevalence to have a two-fold increase (i.e. $I_i(T_{d,i}) = 2s_i$) is expressed by $\log(2)\frac{T_{g,i}}{(R_{0,i}-1)}$.

2.1.2 Stochastic mobility of individuals between populations. The spread of epidemics between populations results from the travel of infected individuals via the WAN. From a given population i, each individual travels to a directly connected population j at a small time interval Δt with probability $w_{ij}\Delta t = F_{ij}\Delta t/N_i$. Suppose population i is directly connect to multiple populations in the WAN, the numbers of susceptible, infectious and recovered travelers that leave population i through an interval Δt, i.e. $X_i(t)$, $Y_i(t)$ and $Z_i(t)$, are simulated with the following set of multinomial random variables:

$$X_i(t) = Multinomial\left(\lfloor S_i(t)\rfloor, w_{i1}\Delta t, ..., w_{iG}\Delta t\right),$$
$$Y_i(t) = Multinomial\left(\lfloor I_i(t)\rfloor, w_{i1}\Delta t, ..., w_{iG}\Delta t\right),$$
$$Z_i(t) = Multinomial\left(\lfloor R_i(t)\rfloor, w_{i1}\Delta t, ..., w_{iG}\Delta t\right),$$

where G counts the number of populations in the WAN, and $Multinomial(n, p_1, ..., p_G)$ denotes a multinomial random variable with n trials and probabilities $p_1, ..., p_G$ [20]. As such, the number of individuals given a specific disease compartment that that travel from population i to j per time interval (e.g. $X_{ij}(t)$) corresponds to the jth component of the corresponding multinomial random variable (e.g. $X_i(t)$).

Table 1: Parameters of the two-population model in which the epidemic origin population i is only connected to population j.

Parameter	Definition
$I_i(t)$	Disease prevalence (number of infectives) in population i at time t
λ_i	Local epidemic growth rate in the origin population i
s_i	Number of initial infections seeded into the origin population i at time 0
w_{ij}	Daily per capita mobility rate from population i to j
α_{ij}	Adjusted mobility rate $\alpha_{ij} = s_i w_{ij}$
T_{ij}^n	The nth arrival time in population j

2.2 Data-driven global metapopulation simulator.

To validate our analytical framework which will be introduced in section 3, we first develop a global metapopulation epidemic simulator, using the algorithm described in section 2.1. Our simulator contains 2,309 populations and 54,106 direct connections. Its structure is similar to the state-of-the-art simulator GLEAM [22] (but without the effect of local commuting which is less important to study the global spread [3]). To build this simulator, we use the 2015 worldwide flight booking data from the Official Airline Guide (OAG, https://www.oag.com) and the Gridded Population of the World Version 4 (GPWv4) dataset from the Columbia University [7]. The OAG dataset provides all flight booking records from all commercial airlines worldwide during 2015, and the GPWv4 dataset provides the highest resolution census data from the 2010 round of Population and Housing Censuses that were collected from hundreds of national statistics departments and organizations.

Ideally, the metapopulation dynamics described in section 2.1 is best implemented with discrete-event simulation algorithms (e.g. Gillespie algorithm [11]). However, explicitly simulating every event of individual infection, recovery and mobility substantially increases the computational burden, which largely exceeds the power of our high-performance

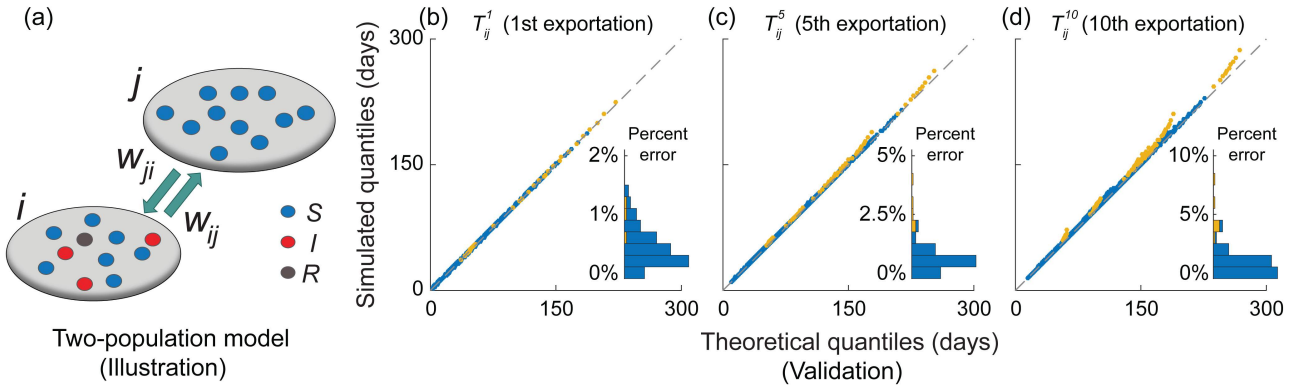


Figure 1: Validating the two-population analytics. (a) Illustration of the two-population model, with the epidemic origin population i only connecting to population j. Table 1 summarizes the parameters. (b)-(c) Q-Q plots for the analytical and simulated quantiles of T_{ij}^{1}, T_{ij}^{5}, and T_{ij}^{10} across 100 epidemic scenarios randomly sampled from the following parameter space using Latin-hypercube sampling: doubling time $T_{d,i}$ and generation time $T_{g,i}$ both between 3 and 30 days, seed size s_i between 1 and 100. Each of the 100 epidemic scenarios is coupled with a set of network parameters randomly sampled with mobility rate w_{ij} between 10^{-6} and 10^{-3} and population size N_i between 0.1 and 10 million, which are chosen according to the OAG and GPWv4 data [25]. Simulated quantiles for each of the 100 scenarios are compiled using 10,000 stochastic realizations. In the Q-Q plots, if data points coincide with the diagonal, the arrival times in the analytical framework are essentially the same as that in the simulation. Data points are colored in blue if the number of exportations X_{ij} is n or above with probability 1 (i.e. $P(X_{ij} \geq n) = 1$), and yellow otherwise. Insets show the corresponding histograms of percent error in $E[T_{ij}^{n}]$.

computing resources. To facilitate the stochastic computing of our global metapopulation epidemic simulator, we use a discrete-time algorithm in which the intra-population epidemic dynamics (see section 2.1.1) and inter-population mobility of travelers (see section 2.1.2) are sequentially simulated for each small time interval Δt. Throughout this work, we set $\Delta t = 0.05$ days, which is sufficiently small to ensure the accuracy of discrete-time simulations [30].

3 ANALYTICAL FRAMEWORK

We formulate the framework by analytically characterizing the probability distribution of EATs for all populations in three metapopulation models with increasingly complex network structure: (i) the simplest two-population model; (ii) the shortest-path-tree of the WAN (WAN-SPT hereafter); and (iii) the whole WAN.

3.1 The two-population model

We start from the two-population model in which the origin population i is only connected to population j (see **Fig. 1a** and **Table 1** for model structure and parameters). This simple model corresponds to the initial stage of a pandemic with infections localized at the origin population (i.e. all the other populations can be merged as a single population that is unaffected to the disease [2]). Our analytical framework grounds on the following two key assumptions [10, 23, 25]:

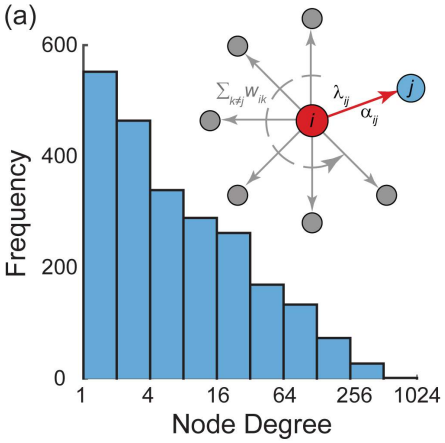
(1) Exportation of infections from population i to j is a nonhomogeneous Poisson process (NPP) [20] with intensity function $w_{ij}I_i(t)$, i.e. the expected number of infections exported from population i to j at time t.
(2) After the epidemic has established in the origin population i, the first few exportations from population i to j occur while disease prevalence is still growing exponentially in the origin i, i.e. $I_i(t) = s_i \exp(\lambda_i t)$.

Under these assumptions, the probability density function (pdf) of T_{ij}^{n} can be expressed in closed-form:

$$f_n(t|\lambda_i, \alpha_{ij}) = \left(\frac{\exp(\lambda_i t) - 1}{\lambda_i}\right)^{n-1} \frac{{\alpha_{ij}}^n}{(n-1)!} \exp\left[\lambda_i t - \frac{\alpha_{ij}}{\lambda_i}(\exp(\lambda_i t) - 1)\right], \quad (1)$$

where $\alpha_{ij} = s_i w_{ij}$ is termed as adjusted mobility rate. To validate this two-population analytics, we compare the analytical and simulated arrival times for a wide range of epidemic scenarios (e.g. the doubling time and generation time both between 3 and 30 days), which are eligible to describe emerging epidemics ranging from pandemic influenza (with doubling time around 4-5 days) to Ebola (with doubling time longer than 20 days). **Figs. 1**(b)-(d) show that Eq. (1)

(a)



(b)

Population (IATA)	Node degree †	Outbound air traffic‡ $F_i = \sum_j F_{ij}$	Outbound mobility rate§ $w_i = \sum_j w_{ij}$
Hong Kong (HKG)	255	60,887	0.0084
Paris (PAR)	455	87,461	0.0052
London (LON)	513	153,817	0.0076
Chicago (CHI)	354	69,618	0.0068
Beijing (BJS)	366	115,042	0.0033
New York (NYC)	444	127,062	0.0094
Mexico City (MEX)	185	36,631	0.0012
Atlanta (ATL)	263	40,407	0.0059
Kuala Lumpur (KUL)	190	52,631	0.0057
Tokyo (TYO)	250	112,239	0.0035
Rio de Janeiro (RIO)	130	33,806	0.0023

† Node degree = number of populations directly connected to the hub population
‡ Outbound air traffic = daily total number of passengers leaving the hub population
§ Outbound mobility rate = rate for each person to leave the hub population per day

Figure 2: Network properties of the hub populations. (a) Histogram shows the distribution of node degree for all populations in the WAN. The node degree of a given population counts the number of populations that are directly connected to that population. Inset illustrates the structure of a travel hub, in which the hub population i is connected to multiple populations, one of which is population j. (b) Illustration of several major hubs in different continents by reporting their node degree, daily outbound traffic volume, and daily outbound per capita mobility rate.

accurately characterizes the arrival times T_{ij}^n for n up to 10 (i.e. the 10th exportation). With Eq. (1), we have the following corollaries:

(1) Exportation of the first n infections is essentially an NPP with intensity function $\alpha_{ij}\exp(\lambda_i t)$.

(2) The cumulative distribution function (cdf) of the nth arrival time is given by

$$F_n\left(t|\lambda_i,\alpha_{ij}\right)=\Gamma\left[n,\frac{\alpha_{ij}}{\lambda_i}\left(\exp\left(\lambda_i t\right)-1\right)\right], \tag{2}$$

where Γ is the lower incomplete Gamma function.

(3) The expected EAT is given by

$$E\left[T_{ij}^1\right]=\frac{1}{\lambda_i}\exp\left(\frac{\alpha_{ij}}{\lambda_i}\right)\mathrm{E}_1\left(\frac{\alpha_{ij}}{\lambda_i}\right), \tag{3}$$

where $\mathrm{E}_m(x)=x^{m-1}\int_x^\infty\left[\frac{\exp(-u)}{u^m}\right]du$ is the exponential integral.

(4) If $\alpha_{ij}<<\lambda_i$ and γ denotes the Euler constant, the expected EAT can be approximated as

$$E\left[T_{ij}^1\right]\approx\frac{1}{\lambda_i}\left[\ln\left(\frac{\lambda_i}{\alpha_{ij}}\right)-\gamma\right], \tag{4}$$

which is congruent with the EAT statistic in Gautreau et al. for estimating the order of epidemic arrival across different populations [10].

(5) The expected time of the nth arrival is given by

$$E[T_{ij}^n]=\frac{1}{\lambda_i}\exp\left(\frac{\alpha_{ij}}{\lambda_i}\right)\sum_{m=1}^{n}\mathrm{E}_m\left(\frac{\alpha_{ij}}{\lambda_i}\right). \tag{5}$$

(6) For any positive integers m and n $(m<n)$, the pdf of T_{ij}^n - T_{ij}^m conditional on T_{ij}^m is simply

$$f_{n-m}\left(t|\lambda_i,\alpha_{ij}\exp\left(\lambda_i T_{ij}^m\right)\right) \tag{6}$$

which corresponds to the time of the $(n-m)$th exportation for an epidemic with seed size $s_i\exp\left(\lambda_i T_{ij}^m\right)$. Using this relation recursively, we deduce that the joint pdf of $T_{ij}^1=t_1,...,T_{ij}^n=t_n$ is simply

$$\prod_{m=1}^{n}f_1\left(t_m|\lambda_i,\alpha_{ij}\exp\left(\lambda_i t_{m-1}\right)\right) \tag{7}$$

for all $0=t_0<t_1<t_2<...<t_{n-1}<t_n$.

(7) The expected time of the $(n-1)$th exportation given an epidemic that starts at time T_{ij}^1 with seed size $s_i\exp\left(\lambda_i T_{ij}^1\right)$ is given by

$$E\left[T_{ij}^n|T_{ij}^1\right]=T_{ij}^1+\frac{1}{\lambda_i}\exp\left(\frac{\alpha_{ij}\exp\left(\lambda_i T_{ij}^1\right)}{\lambda_i}\right)\sum_{m=1}^{n-1}\mathrm{E}_m\left(\frac{\alpha_{ij}\exp\left(\lambda_i T_{ij}^1\right)}{\lambda_i}\right) \tag{8}$$

These corollaries are essential for extending our framework to the WAN-SPT and WAN analysis (see the following two sections).

3.2 The shortest-path tree of the WAN

The WAN-SPT is the dominant sub-network (or backbone) of the WAN, in which each downstream population connects to the epidemic origin via only one path. Brockmann et al. [4, 15] suggested that the epidemic spreads from the origin population to the other populations in the WAN through the WAN-SPT, such that global spread of epidemics through the WAN is primarily driven by the WAN-SPT. We will show that for each population k in the WAN-SPT, the nth arrival time T_{ik}^n can be accurately characterized by the two-population analytics of **Eq. (1)**, where the local epidemic growth rate and adjusted mobility rate are specifically parameterized to account for the hub effect (see section 3.2.1) and continuous seeding effect (see section 3.2.2).

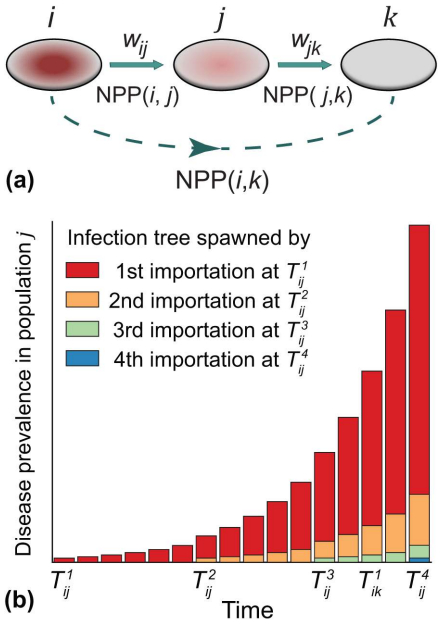


Figure 3: Effect of continuous seeding. (a) Illustration of the epidemic arrival process through an acyclic path that connects the epidemic origin i to population k via population j (i.e. $\psi : i \to j \to k$). (b) In this example, the epidemic arrives at population k after population j has imported three infections from the epidemic origin, i.e. $T_{ij}^3 < T_{ik}^1 < T_{ij}^4$. In the absence of continuous seeding adjustment, infection trees spawned by the second and subsequent importations in population j are ignored [10, 25].

3.2.1 Hub effect. Travel hubs such as Hong Kong, London and Paris have direct flights to multiple populations in the WAN (i.e. their node degree > 1, see **Fig. 2** for illustrations). Given a hub population i, the growth of local disease prevalence $I_i(t)$ can be substantially decreased if a significant proportion of infections travel outward as the epidemic unfolds. To extend our framework to deal with hub populations, we account for the reduction in local disease prevalence by wiping off the hub-effect from local epidemic growth rate λ_i.

Suppose a hub population i is directly connected to two or more populations, one of which is population j (see **Fig. 2(a)**). From the perspective of case arrival process for population j, disease prevalence in population i grows exponentially at rate $\lambda_{ij} = \lambda_i - \sum_{k \neq j} w_{ik}$. Therefore, the pdf of the nth arrival time for population j can be estimated with hub-adjusted two-population analytics $f_n(t|\lambda_{ij}, \alpha_{ij})$, in which infections are exported from population i to j at a rate $w_{ij} I_i(t)$ and disease prevalence in hub population i grows exponentially at the effective growth rate λ_{ij}. Using the hub structure of Hong Kong as an example, **Fig. 4(a)** show that hub-adjusted two-population analytics accurately characterizes the probability distribution of T_{ij}^n for all populations that are directly connected to Hong Kong.

3.2.2 Continuous seeding. Unlike the epidemic origin population which has a single seeding event at time 0, all the other populations in the WAN-SPT can be continuously seeded by infections coming from their upstream populations (illustrated in **Fig. 3**), as exemplified by recent multiple case importations of Zika Virus in Florida that come from the Caribbean [13].

Let D_c be the set of populations that are c degrees of separation from the epidemic origin in the WAN-SPT. Suppose a population k in D_2 is connected to the epidemic origin via population j along the path $\psi : i \to j \to k$. After the epidemic has arrived at population j at time T_{ij}^1, population i continues to export infections to population j before the epidemic arrives at population k at time T_{ik}^1 (illustrated in Fig. 3). According to the two-population model, each imported infection in population j (arriving at times $T_{ij}^1, T_{ij}^2, ...$) spawns an infection tree that grows exponentially at the hub-adjusted rate λ_{jk}. Therefore, the overall disease prevalence in population j, namely $I_j(t)$, is simply the sum of disease prevalence for all these infection trees:

$$I_j(t) = \sum_{m=1}^{\infty} \mathbf{I}\left\{t > T_{ij}^m\right\} \exp\left(\lambda_{jk}\left(t - T_{ij}^m\right)\right)$$

where T_{ij}^m is the mth arrival time in population j, and $\mathbf{I}\{\cdot\}$ is the indicator function. Based on the two-population model, the exportation of infections from population j to k is an NPP with intensity function $w_{jk} I_j(t)$, which is itself a stochastic process because of its dependence on the random variables $T_{ij}^1, T_{ij}^2, ...$ As such, conditional on $I_j(t)$ and hence $T_{ij}^1, T_{ij}^2, ...$, the pdf of T_{ik}^n is

$$g_n\left(t|w_{jk}I_j\right) = f_{Poisson}\left(n-1, w_{jk}\int_0^t I_j(u)\,du\right) w_{jk} I_j(u)$$

for $n = 1, 2....$ The unconditional pdf of T_{ik}^n is thus

$$E_{T_{ij}^1, T_{ij}^2, ...}\left[g_n\left(t|w_{jk}I_j\right)\right]$$

which integrates over the joint pdf of $\left(T_{ij}^1 = t_1, T_{ij}^2 = t_2, ...\right)$.

We conjecture that this highly complex stochastic process can be substantially simplified with little loss of accuracy by using the following assumption: Conditional on T_{ij}^1 (i.e. the EAT for population j), $T_{ij}^m \approx E\left[T_{ij}^m | T_{ij}^1\right]$ for all $m > 1$ (see Eq. 8). Therefore, conditional on T_{ij}^1, we approximate $I_j(t)$ with the following certainty equivalent approximation (CEA):

$$I_j^{CEA}(t) = \sum_{m=1}^{\infty} \mathbf{I}\left\{t > E\left[T_{ij}^m | T_{ij}^1\right]\right\} \exp\left(\lambda_{jk}\left(t - E\left[T_{ij}^m | T_{ij}^1\right]\right)\right)$$
$$= \exp\left(\lambda_{jk}\left(t - T_{ij}^1\right)\right) \sum_{m=1}^{\infty} \mathbf{I}\left\{t > T_{ij}^1 + \Delta T_{ij}^m\right\} \exp\left(-\lambda_{jk}\Delta T_{ij}^m\right)$$

where

$$\Delta T_{ij}^m = E\left[T_{ij}^m | T_{ij}^1\right] - T_{ij}^1$$
$$= \frac{1}{\lambda_{ij}} \exp\left(\frac{\alpha_{ij}\exp\left(\lambda_{ij}T_{ij}^1\right)}{\lambda_{ij}}\right) \sum_{q=1}^{m-1} \mathrm{E_q}\left(\frac{\alpha_{\mathrm{ij}}\exp\left(\lambda_{\mathrm{ij}}\mathrm{T}_{\mathrm{ij}}^1\right)}{\lambda_{\mathrm{ij}}}\right)$$

(see Eq. (8)). The resulting unconditional pdf of T_{ik}^n is simply $E_{T_{ij}^1}\left[g_n\left(t|w_{jk}I_j^{CEA}\right)\right]$ where the pdf of T_{ij}^1 is $f_1\left(\cdot|\lambda_{ij}, \alpha_{ij}\right)$ (see Eq. (1)).

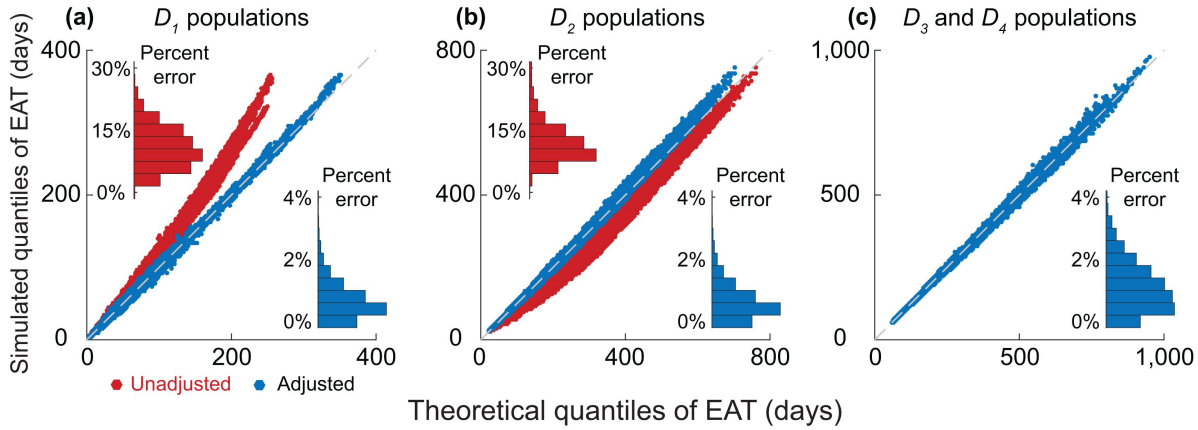


Figure 4: Validating the analytical framework for the WAN-SPT with Hong Kong as the epidemic origin (WAN-SPT-HK). (a)-(c) Q-Q plots for the analytical and simulated quantiles of EATs for all populations in the WAN-SPT-HK across the 100 epidemic scenarios used in Fig. 1. Insets show the corresponding histograms of percent error in expected EAT. (a) EATs for all populations in D_1 before (red) and after (blue) adjusting for the hub-effect. (b) EATs for all populations in D_2 before (red) and after (blue) adjusting for the continuous seeding and path reduction, in which the hub-effect has been adjusted for the epidemic origin and all populations in D_1. (c) EATs for the remaining populations in D_3 and D_4 after adjusting for the hub-effect, continuous seeding and path reduction.

Furthermore, this pdf can in turn be well approximated with $f_n(t|\lambda_\psi, \alpha_\psi)$ where $\lambda_\psi, \alpha_\psi$ are obtained by minimizing the relative entropy [21, 25] for $n = 1$ (i.e. the first exportation). This indicates that the spread of epidemics from the origin to any population in D_2 can be regarded as a two-population model, in which the adjusted mobility rate is α_ψ and the epidemic in the origin grows exponentially at rate λ_ψ. We term this procedure *path reduction*.

Next, consider a longer path $\varphi : i \to j \to k \to m$, i.e. $m \in D_3$. Using path reduction, we first approximate the entire path φ with $\varphi' : i \to k \to m$ where the adjusted mobility rate and adjusted epidemic growth rate in the origin for the connection $i \to k$ are $\lambda_\psi, \alpha_\psi$, respectively. The arrival times of infections for population $m \in D_3$ (i.e. $T^n_{im}, n = 1, 2, ...$) can be estimated using the methods that we have developed for D_2 populations. **Fig. 4** show that recursively using adjustments for the hub-effect and continuous seeding accurately characterizes the arrival times for all populations in the WAN-SPT.

3.3 The whole WAN

The accuracy of our WAN-SPT analysis provides a key insight: for each acyclic path ψ that connects any given population k to the epidemic origin, the epidemic arrival process for population k along this path is well approximated as an *NPP* with intensity function $\alpha_\psi \exp(\lambda_\psi t)$. In the whole WAN, each population might be connected to the epidemic origin via multiple paths, some of which might be intersected and therefore dependent (see Fig. 5(a)). We conjecture that the dependence among such paths is sufficiently weak, such that the overall epidemic arrival process for any population k in the WAN can be characterized with the following method: (i) decomposing all paths that connects the epidemic origin to population k into a set Ψ_{ik} of independent acyclic paths; and then (ii) approximating the EAT for population k by the superposition of the NPPs [20] that correspond to these pseudo-independent paths. Mathematically, the epidemic arrival process for population k is well approximated by an *NPP* with intensity function $\sum_{\psi \in \Psi_{ik}} \alpha_\psi \exp(\lambda_\psi t)$. **Fig. 5** validates that our analytical framework (i.e. synthesis of the two-population analytics, adjustment for the hub-effect, adjustment for continuous seeding, path reduction and path superposition) is accurate for characterizing the EATs for almost all populations in the WAN. The results are robust for all tested 100 epidemic scenarios.

4 CONCLUSIONS

In summary, we have developed an analytical framework that grounds on the basic principles in infectious disease epidemiology and network theory for understanding the dynamics underlying global spread of emerging epidemics. Not only can our framework provides analytical and computational advancement for forecasting EATs for all populations in the WAN, but it also elucidates the dependence of EATs on the epidemiologic parameters (growth rate and seed size) and the network properties of the WAN (air traffic volume and connectivity). Because our framework provides closed-form probability distributions (Eq. (1)), it can also support likelihood-based inference of key epidemiologic parameters from surveillance data on local disease incidence and global case exportations [25]. Ongoing studies deserve to extend the framework to account for more complex factors including the

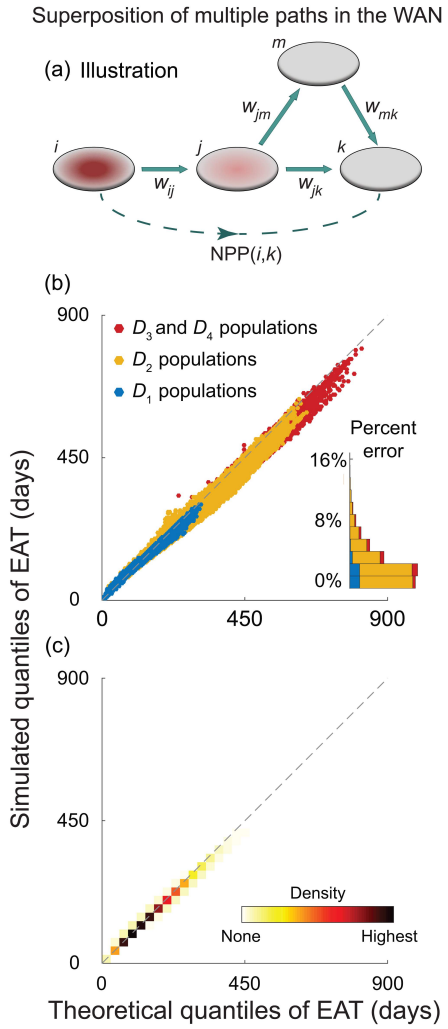


Figure 5: Validating the analytical framework for the WAN. The epidemic origin is Hong Kong as in Fig. 4. (a) Q-Q plots for the analytical and simulated quantiles of EATs for all populations in the WAN. Analytical EATs are computed using the NPP superposition as described in the section 3.3, while simulated EATs are generated from our global metapopulation simulator as described in the section 2.2. Data points are colored in blue for D_1 populations, yellow for D_2 populations, and red for D_3 and D_4 populations. (b) Density of the data points in (a) to show that nearly all the 230,800 Q-Q plots coincide with the diagonal, which demonstrates the congruence between analytical and simulated EATs.

stochasticity of intra-population transmission dynamics [29] and seasonal travel patterns [8, 27].

A APPENDICES

A.1 SE_mI_nR model for epidemic spreading within each population

In the main text, we build the analytical framework using the SIR model within each population. Here we extend the theory to SE_mI_nR models [26] in which:

(1) The duration of latency is gamma distributed with mean D_E and m subclasses (i.e. with shape m and rate $b_E = m/D_E$.
(2) The duration of infectiousness is gamma distributed with mean D_I and n subclasses (i.e. with shape n and rate $b_I = n/D_I$.

For any given population, let $S(t), R(t)$ be the number of susceptible and recovered individuals, respectively, $E_i(t)$ the number of individuals in the ith latent subclass, and $I_j(t)$ the number of individuals in the jth infectious subclass. The SE_mI_nR model is described by the following differential equations:

$$\frac{dS(t)}{dt} = -\beta \frac{S(t)}{N} \sum_{j=1}^{n} I_j(t)$$

$$\frac{dE_1(t)}{dt} = \beta \frac{S(t)}{N} \sum_{j=1}^{n} I_j(t) - b_E E_1(t)$$

$$\frac{dE_i(t)}{dt} = b_E \left(E_{i-1}(t) - E_i(t)\right) \quad \text{for} \quad i = 2, ..., m$$

$$\frac{dI_1(t)}{dt} = b_E E_m(t) - b_I I_1(t)$$

$$\frac{dI_j(t)}{dt} = b_I \left(I_{j-1}(t) - I_j(t)\right) \quad \text{for} \quad j = 2, ..., n$$

$$\frac{dR(t)}{dt} = b_I I_j(t).$$

During the early stage of the epidemic (such that $S(t) \approx N$), the prevalence of latent and infectious people both grows exponentially at rate λ, which is the solution to the following equation [26]:

$$\lambda\left(\lambda + \frac{m}{D_E}\right)^m - \beta\left(\frac{m}{D_E}\right)^m \left(1 - \left(\frac{\lambda D_I}{n} + 1\right)^{-n}\right) = 0$$

That is, the prevalence of latent and infectious individuals are well approximated by $\bar{E}\exp(\lambda t)$ and $\bar{I}\exp(\lambda t)$, respectively, where $\bar{E}$ and $\bar{I}$ depend on the initial conditions and parameters of the differential equation systems (the analytical expressions of $\bar{E}$ and $\bar{I}$ are obtained by solving the linearized system with $S(t) = N$). If a proportion $1 - p_E$ and $1 - p_I$ of the latent and infectious people refrain from air travel because of their infections, the seed size s_0 in the main text is simply $p_E\bar{E} + p_I\bar{I}$.

ACKNOWLEDGMENTS

We thank M. Lipsitch, J.M. Read, B.J. Cowling, P. Wu, K. Leung, H. Choi, N. Leung, S. Ali, J. Wong, V.J. Fang, Z. Wang, L. Chen, Y. Zhang and Y. Lin for helpful discussions. We thank C.K. Lam for assistance in data processing and technical support. We thank the Official Airline Guide and

Center for International Earth Science Information Network at Columbia University for the assembly of databases. This research was conducted in part using the research computing facilities and advisory services offered by Information Technology Services, The University of Hong Kong; and was done in part on the Olympus High Performance Compute Cluster at the Pittsburgh Supercomputing Center at Carnegie Mellon University, which is supported by National Institute of General Medical Sciences MIDAS Informatics Services Group under Grant No.: 1U24GM110707. This research was supported by Harvard Center for Communicable Disease Dynamics from the National Institute of General Medical Sciences MIDAS Initiative under Grant No.: U54GM088558, Research Grants Council Collaborative Research Fund under Grant No.: CityU8/CRF/12G, and two commissioned grants from the Health and Medical Research Fund from the Government of the Hong Kong SAR under Grant No.: HKS-15-E03, HKS-17-E13. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences and the National Institutes of Health.

REFERENCES

[1] Roy M. Anderson and Robert M. May. 1991. *Infectious Diseases of Humans: Dynamics and Control* (1st ed.). Oxford Univ. Press, Oxford, UK.

[2] Paolo Bajardi, Chiara Poletto, Jose J Ramasco, Michele Tizzoni, Vittoria Colizza, and Alessandro Vespignani. 2011. Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. *PloS one* 6, 1 (Jan. 2011), e16591. https://doi.org/10.1371/journal.pone.0016591

[3] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences* 106, 51 (Dec. 2009), 21484–21489. https://doi.org/10.1073/pnas.0906910106

[4] Dirk Brockmann and Dirk Helbing. 2013. The hidden geometry of complex, network-driven contagion phenomena. *Science* 342, 6164 (Dec. 2013), 1337–1342. https://doi.org/10.1126/science.1245200

[5] Dennis Carroll, Peter Daszak, Nathan D Wolfe, George F Gao, Carlos M Morel, Subhash Morzaria, Ariel Pablos-Méndez, Oyewale Tomori, and Jonna AK" Mazet. 2018. The global virome project. *Science* 359, 6378 (Feb. 2018), 872–874. https://doi.org/10.1126/science.aap7463

[6] Benjamin J Cowling, Dennis KM Ip, Vicky J Fang, Piyarat Suntarattiwong, Sonja J Olsen, Jens Levy, Timothy M Uyeki, Gabriel M Leung, JS Malik Peiris, Tawee Chotpitayasunondh, et al. 2013. Aerosol transmission is an important mode of influenza A virus spread. *Nature communications* 4 (June 2013), 1935. https://doi.org/10.1038/ncomms2922

[7] Erin Doxsey-Whitfield, Kytt MacManus, Susana B Adamo, Linda Pistolesi, John Squires, Olena Borkovska, and Sandra R Baptista. 2015. Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4. *Papers in Applied Geography* 1, 3 (July 2015), 226–234. https://doi.org/10.1080/23754931.2015.1014272

[8] Anne Ewing, Elizabeth C Lee, Cécile Viboud, and Shweta Bansal. 2016. Contact, travel, and transmission: The impact of winter holidays on influenza dynamics in the United States. *The Journal of infectious diseases* 215, 5 (Dec. 2016), 732–739. https://doi.org/10.1093/infdis/jiw642

[9] J Patrick Fitch. 2015. Engineering a global response to infectious diseases. *Proc. IEEE* 103, 2 (March 2015), 263–272. https://doi.org/10.1109/JPROC.2015.2389146

[10] Aurélien Gautreau, Alain Barrat, and Marc Barthelemy. 2008. Global disease spread: statistics and estimation of arrival times. *Journal of theoretical biology* 251, 3 (April 2008), 509–522. https://doi.org/10.1016/j.jtbi.2007.12.001

[11] Daniel T Gillespie. 1977. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry* 81, 25 (Dec. 1977), 2340–2361. https://doi.org/10.1021/j100540a008

[12] Bryan Grenfell and John Harwood. 1997. (Meta) population dynamics of infectious diseases. *Trends in ecology & evolution* 12, 10 (July 1997), 395–399. https://doi.org/10.1016/S0169-5347(97)01174-9

[13] Nathan D Grubaugh, Jason T Ladner, Moritz UG Kraemer, Gytis Dudas, Amanda L Tan, Karthik Gangavarapu, Michael R Wiley, Stephen White, Julien Thézé, Diogo M Magnani, et al. 2017. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* 546, 7658 (jun 2017), 401–405. https://doi.org/10.1038/nature22400

[14] Edward C Holmes, Andrew Rambaut, and Kristian G Andersen. 2018. Pandemics: spend on surveillance, not prediction. *Nature* 558 (June 2018), 180–182. https://doi.org/10.1038/d41586-018-05373-w

[15] Flavio Iannelli, Andreas Koher, Dirk Brockmann, Philipp Hövel, and Igor M Sokolov. 2017. Effective distances for epidemics spreading on complex networks. *Physical Review E* 95, 1 (Jan. 2017), 012313. https://doi.org/10.1103/PhysRevE.95.012313

[16] Matt J. Keeling and Pejman Rohani. 2007. *Modeling Infectious Diseases in Humans and Animalss*. Princeton Univ. Press, Princeton, NJ.

[17] Madhav Marathe and Anil Kumar S Vullikanti. 2013. Computational epidemiology. *Commun. ACM* 56, 7 (July 2013), 88–96. https://doi.org/10.1145/2483852.2483871

[18] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, et al. 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine* 5, 3 (March 2008), e74. https://doi.org/10.1371/journal.pmed.0050074

[19] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. 2015. Epidemic processes in complex networks. *Reviews of modern physics* 87, 3 (Aug. 2015), 925. https://doi.org/10.1103/RevModPhys.87.925

[20] Sheldon M. Ross. 1996. *Stochastic Processes* (2nd. ed.). John Wiley & Sons, New York, NY.

[21] Joy A. Thomas and Thomas M. Cover. 2006. *Elements of Information Theory* (2nd. ed.). John Wiley & Sons, Hoboken, NJ.

[22] Michele Tizzoni, Paolo Bajardi, Chiara Poletto, José J Ramasco, Duygu Balcan, Bruno Gonçalves, Nicola Perra, Vittoria Colizza, and Alessandro Vespignani. 2012. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC medicine* 10, 1 (Dec. 2012), 165. https://doi.org/10.1186/1741-7015-10-165

[23] Gianpaolo Scalia Tomba and Jacco Wallinga. 2008. A simple explanation for the low impact of border control as a countermeasure to the spread of an infectious disease. *Mathematical biosciences* 214, 1-2 (July 2008), 70–72. https://doi.org/10.1016/j.mbs.2008.02.009

[24] Lin Wang and Xiang Li. 2014. Spatial epidemiology of networked metapopulation: An overview. *Chinese Science Bulletin* 59, 28 (Oct. 2014), 3511–3522. https://doi.org/10.1007/s11434-014-0499-8

[25] Lin Wang and Joseph T Wu. 2018. Characterizing the dynamics underlying global spread of epidemics. *Nature Communications* 9, 1 (Jan. 2018), 218. https://doi.org/10.1038/s41467-017-02344-z

[26] Helen J Wearing, Pejman Rohani, and Matt J Keeling. 2005. Appropriate models for the management of infectious diseases. *PLoS Medicine* 2, 7 (Jul 2005), e174. https://doi.org/10.1371/journal.pmed.0020174

[27] Amy Wesolowski, Elisabeth zu Erbach-Schoenberg, Andrew J Tatem, Christopher Lourenço, Cecile Viboud, Vivek Charu, Nathan Eagle, Kenth Engø-Monsen, Taimur Qureshi, Caroline O Buckee, et al. 2017. Multinational patterns of seasonal asymmetry in human movement influence infectious disease dynamics. *Nature communications* 8, 1 (Dec. 2017), 2069. https://doi.org/10.1038/s41467-017-02064-4

[28] World Health Organization (WHO). 2018. 2018 Annual review of diseases prioritized under the Research and Development Blueprint. Retrieved July 8, 2018 from http://www.who.int/blueprint/priority-diseases/en/

[29] Joseph T. Wu and Benjamin J. Cowling. 2011. The use of mathematical models to inform influenza pandemic preparedness and response. *Experimental Biology and Medicine* 236, 8 (Aug. 2011),

955–961. https://doi.org/10.1258/ebm.2010.010271
[30] Joseph T Wu, Gabriel M Leung, Marc Lipsitch, Ben S Cooper, and Steven Riley. 2009. Hedging against antiviral resistance during the next influenza pandemic using small stockpiles of an alternative chemotherapy. *PLoS medicine* 6, 5 (May 2009), e1000085. https://doi.org/10.1371/journal.pmed.1000085
[31] Xiaoxu Wu, Yongmei Lu, Sen Zhou, Lifan Chen, and Bing Xu. 2016. Impact of climate change on human infectious diseases: Empirical evidence and human adaptation. *Environment international* 86 (Jan. 2016), 14–23. https://doi.org/10.1016/j.envint.2015.09.007

Cleanliness Campaign V/S Sanitation Related Diseases - Are they parallel in public perspective?

Aarzoo Dhiman
Indian Institute of Technology Roorkee,
Uttarakhand- 247667, India
aarzoodhiman.dcs2017@iitr.ac.in

Durga Toshniwal
Indian Institute of Technology Roorkee,
Uttarakhand- 247667, India
durgafec@iitr.ac.in

Soumya Somani
Symbiosis Institute of Technology,
Pune, Maharashtra- 412115, India
soumya.somani@sitpune.edu.in

Preeti Malik
Indian Institute of Technology Roorkee,
Uttarakhand- 247667, India
parimalik.pcs2016@iitr.ac.in

ABSTRACT

Social media data is playing an important role in healthcare and it is being used for performing many epidemiological tasks such as outbreak surveillance, intervention surveillance, modeling the disease spread through a community, etc. This is due to easy and early availability of social data unlike the clinical data sources which have very limited availability. Twitter data, being in the form of micro-blogs, is the most effective way of performing any study on the thought process of the public as people tweet about anything and everything on their handles. In the present research work, the Twitter data related to an Indian National Level cleanliness campaign, called *Swachh Bharat Abhiyan (SBA)* and the diseases which occur due to lack of cleanliness such as Dengue, Malaria, Diarrhoea, etc. has been collected for the period of 1 January, 2018 to 31 March, 2018. A demographic and temporal analysis of the Twitter data has been performed to compare and contrast the perception of Indian citizens towards SBA and diseases caused by lack of cleanliness. A study of the impact of SBA on occurrence of many diseases which occur due to lack of cleanliness has also been performed. Our experiments showed that the tweets related to both the topics were not very correlated and sentiment analysis of such tweets showed that most of tweets had neutral sentiments.

KEYWORDS

Data Mining, Cleanliness Campaign, Common Water, Sanitation, Swachh Bharat Abhiyan

ACM Reference Format:
Aarzoo Dhiman, Durga Toshniwal, Soumya Somani, and Preeti Malik. 2018. Cleanliness Campaign V/S Sanitation Related Diseases - Are they parallel in public perspective?. In *Proceedings of epiDAMIK.* epiDAMIK, London, UK, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

epiDAMIK, Workshop, London, UK. August 20, 2018

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Health related issues can be caused due to many reasons for example, bacterial, viral, fungi, microbial, genetic, parasitic etc. Broadly, the sources of all these reasons can be environmental, social, economic, physical, chemical, political and biological factors etc. Epidemiologists carry out investigations that examine all the above mentioned socio-economic, political and environmental factors that cause health related issues to improve public health. This study is useful in improving the overall public health and health of the disadvantaged, find out the relation between genetic factors, environmental factors and personal behaviors and their interplay, diminish the sources of disease causing agents and study the influence and effects of health programs and services on overall public health.

Multiple air-borne, water-borne and food-borne communicable diseases such as Diarrhoea, Dengue and Typhoid etc. are caused by the lack of proper sanitation, solid and liquid waste management and cleanliness. The occurrence of these diseases can create outbreaks in a few days. That's why, epidemiologists have started working on *Early outbreak detection systems* [2], which are able to detect the outbreaks in the earlier stages of its spread. The primary mode of doing this is by modeling the spread of an outbreak and then predicting the future number of cases by using some available data sources. Traditional data sources have their roots in collecting data through in-patient records of several clinics and hospitals in different regions of the country. However, one primary drawback of these data sources is the delay in availability of data by few weeks or months. Thus, epidemiologists have moved their attention from traditional data sources to web data sources such as social media networks, blogging and micro-blogging networks and search engine query logs. The web data is used because of the ease of access, faster availability and huge amount, which can help in detection of spread of a disease in the earliest stages possible.

Along with monitoring the spread of a disease in a community, epidemiology also deals with providing the measures to eliminate the causes of the spread of the diseases. Such

measures mostly include the preventive measures for example, vaccination, campaigns and teaching people about risks and abuses of drugs etc. These measures of intervention also need to be monitored to track their effects in the community so that appropriate actions can be taken. One such national level cleanliness campaign, called *Swachh Bharat Abhiyan (SBA)* was launched by government of India on October 02, 2014, to improve the cleanliness situation in India. One primary aim of this campaign is to make India free from open defecation and achieve 100 percent scientific solid waste management by October 2019. However, there are very few statistics provided by the government, which can ensure the level of involvement and awareness among people towards SBA and diseases which occur due to lack of cleanliness. In this paper, first a comparison study of geographic and temporal distribution of the tweets related to SBA and sanitation related diseases has been performed. This study will help in determining the awareness of the citizens of India about the causal relationship between the two topics: SBA and common water and sanitation related diseases.

There are several diseases which are caused due to lack of cleanliness such as Dengue, Malaria etc. However, there is very limited availability of the standard clinical data sources which can provide exact number of cases of these diseases. Hence, Twitter data pertaining to these diseases has been collected and studied to study the impact of SBA on prevalence of diseases caused due to lack of cleanliness for the period of 1, January 2018 to 31, March 2018.

There has been much research work done related to monitoring the outbreak surveillance and tracking the impact of any intervention in a community using the social media data. However, there has not been much research work done which aimed to monitor the relationship between the two. Hence, in this paper, Twitter data is used to monitor the awareness of people on relationship between the cleanliness campaign i.e. Swachh Bharat Abhiyan(SBA) and spread of common water and sanitation related diseases such as Malaria, Dengue etc.

The rest of the paper has been organized as follows. Section 2 contains the related work in the field of epidemiology. Section 3 contains our proposed work which includes the data set description and the methodology used for our study. Section 4 contains results and discussion, which highlights important findings of our analysis. Finally, the paper ends with conclusion and references.

2 RELATED WORK

Social media sites have become the source of providing a variety of features which fulfill many purposes such as social networking, professional networking, media sharing content production, knowledge and information aggregation, virtual reality and gaming environment etc.[9], professional education, organizational promotions, patient care, patient education and spreading information about public health programs. Essentially, Social media allows to ask, and answer, questions were never thought to be possible.

In [1], Rumi Chunara et al. performed early epidemiological assessment using various social media sources during the 2010 Haitian cholera outbreak. Their study showed a good correlation among the official data and social media data which was available up to 2 weeks earlier. Through this study, the authors proposed that social media data can be used in replacement of official data in an outbreak setting to get timely estimates of the disease dynamics. Another approach by J. Gomide et al. [4] studied the extent of Twitter as a tool for surveillance of Dengue epidemic. The methodology proposed in the research work was based on four dimensions: volume, location, time and public perception. First, the public perception dimension was explored by performing sentiment analysis, which filtered out the content that is not relevant for Dengue surveillance. Then, the number of cases reported by official statistics and the number of posts on Twitter during the same time period was correlated and verified. The authors exploited the spatio-temporal dimension of the data to create clusters and the quality of the clusters were then compared to the official data. Another recent study by King-Wa Fu et al. [3] aimed to provide the baseline data model for Zika Virus related English tweets. Its motivation came from the 2015-2016 Zika Virus epidemic in The United States. This study focused on ZIKV-infected pregnancy which could be complicated with fetal microcephaly and long-term developmental disability. As stated in the study, "Epidemiological evidences suggested that ZIKV might cause GuillainBarre syndrome. The World Health Organization (WHO) declared it a Public Health Emergency of International Concern (PHEIC) on February 1, 2016". The authors presented an incidence trend analysis of Zika Virus-related Twitter data and content analysis of a cross-sectional sample of Zika Virus-related English Tweets in their research work.

Now and again, the government keeps on introducing preventive measures to eliminate the socio-economic, environmental, chemical and biological factors behind the causes and spread of a disease to improve public health in a community. The effects of these preventive measures need to be tracked so that appropriate actions could be taken. In 2010, Scanfeld et al.[6] examined the data from Twitter to track the misuse and misunderstanding related to the use of antibiotics in the society by using content analysis techniques. Later in 2017, Shah et al.[7] traced the change in behavior of users search data before and after the introduction of Rota Virus and Noro Virus vaccination in US, UK and Mexico by using the data from Google quantified Internet Query Share (IQS). SBA was first launched in 2014 and since then very less research work has been done related to it and there is no research work done which compares the awareness of these two issues in common public. In 2015, Sahil Raj et al. [5] collected tweets related to SBA and performed simple sentiment analysis to find out perception of Indian citizens towards SBA. Later in 2016, Devendra et al. [8] tested their sentiment analysis tool Senti-Meter on Twitter data related to SBA. They studied 1200 tweets collected for the period of January 2016 to March 2016 and performed manual tagging to evaluate the accuracy of their tool. Both of these works worked on very less number

of tweets and did not consider any other demographic details of the Indian cities and states.

3 PROPOSED WORK

The effect of programs like SBA on the occurrence of sanitation related diseases is yet unexplored. To see these changes in the society it's a must that people are aware about the relation between cleanliness and diseases. The primary aim of this study is to track the involvement and perception of people towards SBA and the sanitation and water related diseases. The data related to these topics has been collected separately using the Twitter API and then compared to determine any causal relationships among them.

3.1 Dataset

This section gives some details on the datasets used for the study. The Twitter data related to SBA and the sanitation related diseases has been used to monitor the involvement of people in both these topics and to determine any relationship among them. The datasets have been collected over Twitter Live Stream using Suitable keywords for a period of three months i.e., January 2018 to March 2018.

3.1.1 Disease Data. Tweets for diseases related to common water and sanitation have been collected for a period of three months using Twitter API. Details can be seen in Table 1.

Table 1: Disease Data Description

Sr.No.	Attribute	Values
1	Number of diseases	9
2	Names of diseases	Chikungunya, Cholera, Dengue, Diarrhoea, Hepatitis, Japanese Encephalitis, Malaria, Typhoid, and Zika
3	Tweets Collected	18 thousand

3.1.2 Swachh Bharat Abhiyan Data. SBA related tweets have been accumulated using the keywords elaborated in Table 2. 4 hundred thousand tweets have been collected for the given period with specific numbers given in the Table 2.

3.2 Proposed Methodology

Fig. 1 briefly represents the steps of the proposed method. All these steps are explained in the following subsections.

3.2.1 Data Preprocessing. Twitter users need not specify their locations in the account details. This may be the reason why not all the tweets collected have a location attribute in them. Such tweets are needed to be filtered out for further processing so that the locations could be captured.

3.2.2 Demographic Analysis. Let any state or union territory of India be denoted as S_{i}, where

$$i = 1 \text{ to } n,$$

and any disease be denoted as D_{j}, where

Table 2: Description of Twitter data collected using keywords related to SBA

Hashtags	Examples	Tweet example	Number of Tweets
General	Swachh Bharat Abiyan MyCleanIndia Open Defecation MyCity-MyPride	@marineravin: @tavleen_singh anything you wanna more to add abt swachh Bharat Abhiyan ... @sanjayuvacha @amitmehra	322,287
Toilet Related		@paramiyer_: Congratulations to Team @swachhbharat. Tirunelveli district in Tamil Nadu has been declared #OpenDefecationFree.	26,846
Cities Related	SwachhUP SwachhJhar	RT @lezlietripathy: Participated in Cleaning #Vesave Beach Today. An initiative by @AfrozShah1 Supported by @Dev_Fadnavis @AUThackeray Today. #SwachhBharat #Swachhmaharashtra #swachhversova	24,736
Rural Area Related	ZSBP SbmZSBP	@kishanganjzsbp: Morning follow up and pit digging in Gachpada Panchayat #ZSBP #SwachhBharat #SwachhBihar #SBM-Gramin @SwachhBihar @LSBA_Bihar @swachhbharat	86,868

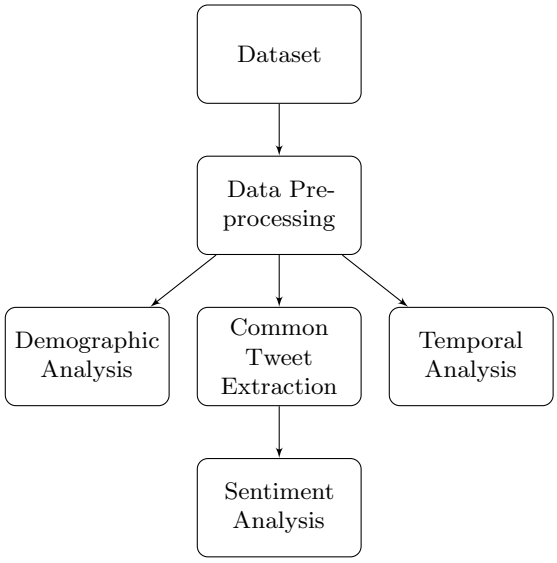


Figure 1: Proposed Methodology

$$j = 1 \text{ to } m,$$

Total number of tweets about diseases is denoted by TD,

$$TD = \sum_{i=1}^{n} TD_i$$

where TD_{i} is defined as,

$$TD_i = \sum_{j=1}^{m} TD_{ij}$$

where,

TD_{ij} = No. of tweets from state i about disease j

Total number of tweets about SBA is denoted by TS,

$$TS = \sum_{i=1}^{n} TS_i$$

where,

$$TS_i = \text{No. of tweets from state i about SBA}$$

TD_{i} and TS_{i} values are compared to give the state wise distribution.

3.2.3 Temporal Analysis. Let the number of weeks be denoted by k, where $k = 1$ to w. The total number of tweets of any disease be denoted by TD_{j}, where

$$TD_j = \sum_{k=1}^{w} TD_{jk}$$

where,

$$TD_{jk} = \text{No. of tweets about a disease j in week k}$$

The total number of tweets of SBA be denoted by TS, where

$$TS = \sum_{k=1}^{w} TS_k$$

where,

$$TS_k = \text{No. of tweets about SBA in week k}$$

Normalized values of TD_{jk} and TS_{k} values are compared to give a weekly distribution.

3.2.4 Common Tweet Extraction. The datasets are further processed to extract the tweets which talk about SBA and any sanitation related disease at the same time. To extract such tweets we performed a simple keyword search for the mention of both the topics from the Twitter data at the same time. This extraction has been done to study the distribution of the parallel thoughts of people on both the topics.

3.2.5 Sentiment Analysis. Sentiment analysis is a supervised classification process to predict the opinion of a person through the text which is related to some topic. We used sentiment analysis on our Twitter corpus to capture the opinions and sentiments of people towards SBA. Each tweet has been classified into three opinions: positive, negative and neutral by using *Word Sense Disambiguation, Senti Word Net and word occurrence statistics using movie review corpus*. We used the dedicated sentiment classification library of python for our study. If sentiment score value comes out to be greater than 0 then the sentiment is classified as positive, if it comes out to be less than 0 then the sentiment is classified as negative and otherwise neutral. Sentiment Analysis is performed on the tweets which talk about both the topics at the same time to find the opinion of people regarding the two topics.

4 RESULTS AND DISCUSSIONS

In this section, we have highlighted some of the important results of our experimentation. Section 4.1 presents the involvement and emotions about SBA and common water & sanitation related diseases in different states in India. Section 4.2 presents the weekly distribution of tweets for the period of three months. Section 4.3 gives the monthly distribution of common tweets among the SBA data and common water & sanitation related disease data to derive the perception of common public about the relationship among them. All the experimentations have been performed using Python.

4.1 Demographic Analysis

First, we present a state level distribution of total tweets for three months period related to SBA and common water & sanitation related diseases as shown in Figure 2. As visible from the figure, the number of SBA tweets is greater than that of the tweets related to common water & sanitation related diseases. This depicts that the overall SBA related public awareness is higher than that of sanitaion related diseases. It also depicts that in case of SBA, Maharashtra has shown the maximum number of tweets and in case of diseases, Delhi has shown the maximum number of tweets. There are some other states as well where number of SBA related tweets is very high but number of disease related tweets are very less e.g. in case of Madhya Pradesh. This can be due to the fact that Madhya Pradesh has been ranked 1st in SBA rankings 2017 given by government of India.

As seen from Figure 2, number of tweets related to SBA are overshadowing the number of tweets related to the diseases. Hence, we extract sanitation related tweets out of the SBA tweets using the 'toilet' related keywords such as 'open defecation' and 'toilet' etc. Figure 3 gives a state level distribution of sanitation related tweets and sanitation related diseases. The figure depicts that the number of disease related tweets is greater than that of sanitation related SBA tweets. As seen from Figure 2 and Figure 3, the difference is number of tweets in all the three types of tweets is very high. There is very less correlation between the number o ftweets for all the three sets, which means that the people who are talking about one topic may not be talking about the other topic at the same time. This shows that although the overall popularity of SBA is more than the awareness of common water & sanitation related diseases but when it comes to specific reasons behind SBA (i.e. improving cleanliness situation), people are not very aware of its relationship with the effects of SBA (i.e. elimination of causes behind sanitation related diseases).

Further, to make the study exhaustive, Pearson's correlation of the normalized number of tweets (i.e. percentage of number of tweets) related to SBA, sanitation and sanitation and water related disease has been performed. Table 3 gives the correlation value and the P-values for the same. The correlation for most of diseases is found to be negative that means they have an inverse relationship with the number of tweets related to SBA. However, the P-values are mostly greater than 0.05, which may be due to low sample size. This correlation is primarily quantitative in nature. Most of the correlation values are found to be negative. This shows that when there are high number of tweets related to SBA and sanitation, there may be less number of tweets related to some diseases such as Chikungunya, Cholera, Zika etc. The positivity in correlation is also found to be near to zero such as in case of Dengue, Hepatitis and Malaria. The reason behind such uncorrelated behavior may be a few number of

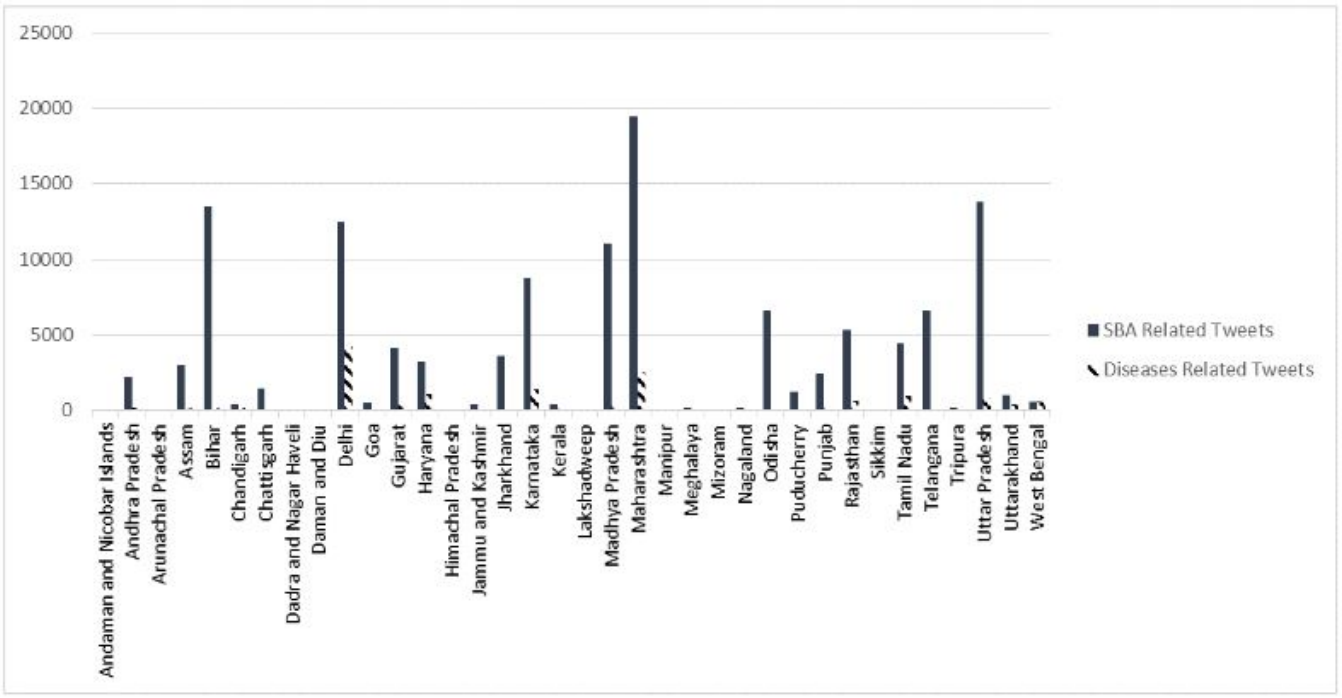


Figure 2: State wise distribution of SBA v/s diseases tweets

tweets but this also signifies that the people do not view both of these topics as correlated.

4.2 Temporal Analysis

The above study gave us the demographic distribution of the tweets related to SBA and sanitation related diseases. Now, to study the change in number of tweets over the given period we have performed weekly trend analysis. Here, the change in percentage of tweets related to these diseases and SBA over the three month period has been analyzed as shown in Figure 4. A large number of tweets related to Dengue, Hepatitis and Malaria are prominent in this duration of the year. Also, the number of tweets related to Dengue and Malaria increases in the month of March which can be due to the increase in mosquitoes in any area. The change in number of tweets related to SBA and sanitation over the three months period can also be seen in the figure.

4.3 Common Tweets Analysis

From the above distribution no real correspondence can be seen between the two topics. To find this, the tweets having both the keywords, from SBA as well as diseases, are found out. As can be seen in Figure 5, the overall number of these tweets which mention SBA and sanitation related disease at the same time are very few. So, people are supporting SBA and talking about health issues individually but there is a lack of awareness among people about how SBA is making a difference in terms of elimination of sanitation and water related diseases. Though the number of tweets are increasing during March because this is the period of occurence of water and sanitation related diseases e.g. Dengue, Diarrhoea, Malaria etc. but these are considerably very low.

4.3.1 Sentiment Analysis. To study the opinions of the people towards both the topics, we extracted the tweets which talk about both the topics i.e. SBA and diseases related to sanitation and performed sentiment analysis on them. There were very few tweets which were talking about both the topics at the same time and the sentiment analysis of these tweets show that most of the tweets show a neutral sentiment. Large number of neutral tweets show that most of the tweets are generally related to spreading awareness about the cleanliness

Table 3: Correlation between the number of tweets related to SBA, sanitation and santitaion and water related diseases

Disease	SBA related tweets	P-value	Santitation related tweets	P-value
Chikungunya	-0.461	0.113226	-0.479	0.097393
Cholera	-0.422	0.150817	-0.516	0.071292
Dengue	0.0175	0.954782	-0.29	0.336281
Diarrhoea	-0.154	0.615615	0.2009	0.510394
Hepatitis	0.0619	0.840839	-0.009	0.977562
Japanese Encephalitis	0.1307	0.670305	-0.046	0.881549
Malaria	0.0109	0.971833	-0.109	0.722413
Typhoid	0.1476	0.630398	-0.026	0.931693
Zika	-0.445	0.127708	-0.579	0.037992

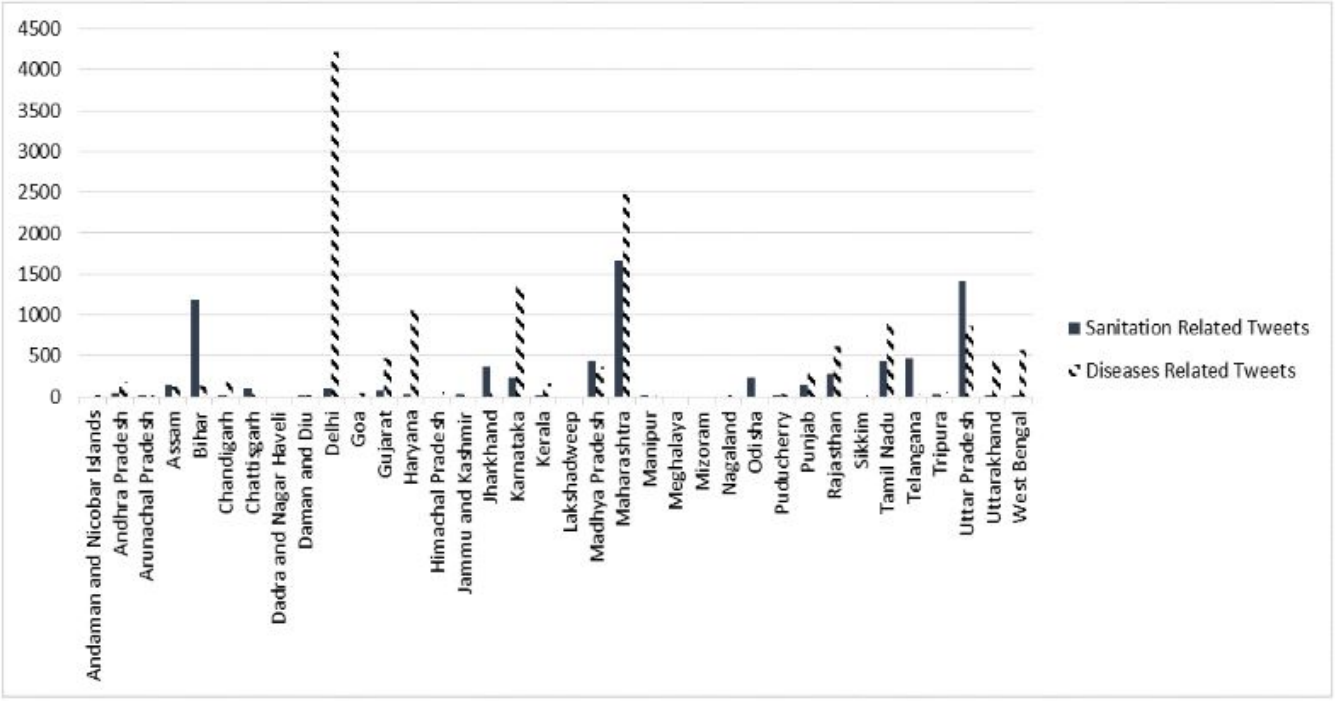


Figure 3: State wise distribution of sanitation related v/s diseases tweets

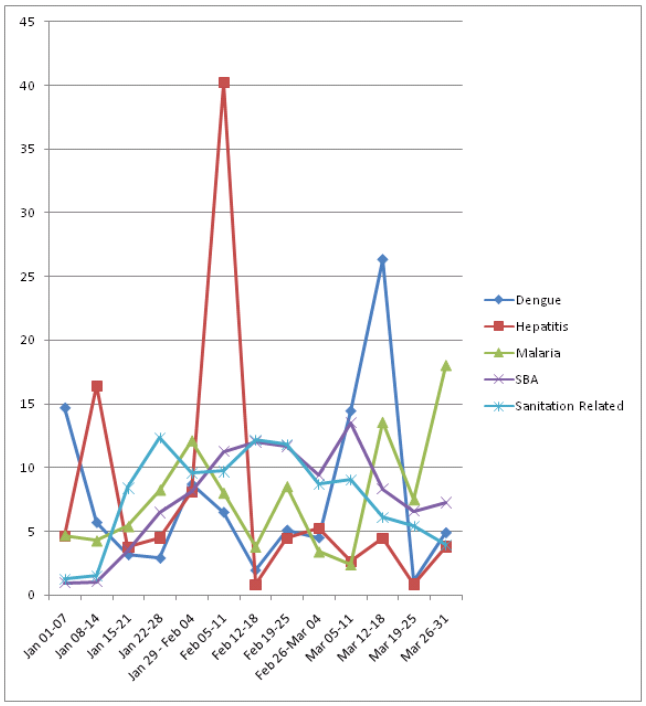


Figure 4: Weekly distribution of percentage of tweets

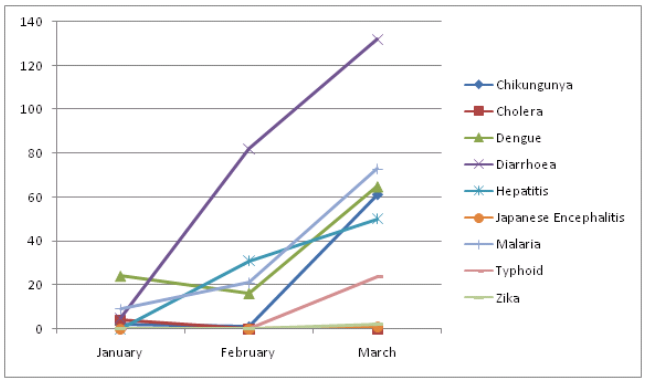


Figure 5: Number of common tweets

campaign and its benefits in context to different diseases that are caused due to lack of cleanliness. This supports our previous deduction that people are aware about both the topics separately, but they are not much interested in talking about both the topics as being related to each other.

5 CONCLUSION

In this paper, the Twitter data is used to capture the insights of public on two topics that have a causal relationship among them i.e. SBA and sanitation related diseases. Through this study, the perception of people about the relationship between these two topics has been monitored. Here, the SBA and disease related data has been analyzed separately as well as

collectively. Cleanliness and diseases are well connected terms for real but the results from this work announce otherwise. Results show that people are generally aware of SBA as well as sanitation related diseases on an individual level but they are very less aware about the relationship between the two. This can also be seen as a negative fact as people are tweeting with popular SBA hash-tags without knowing its value and effects in reducing the disease occurrence.

REFERENCES

[1] Rumi Chunara, Jason R Andrews, and John S Brownstein. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American journal of tropical medicine and hygiene* 86, 1 (2012), 39–45.

[2] Ed De Quincey and Patty Kostkova. 2009. Early warning and outbreak detection using social networking websites: The potential of twitter. In *International Conference on Electronic Healthcare*. Springer, 21–24.

[3] King-Wa Fu, Hai Liang, Nitin Saroha, Zion Tsz Ho Tse, Patrick Ip, and Isaac Chun-Hai Fung. 2016. How people react to Zika virus outbreaks on Twitter? A computational content analysis. *American journal of infection control* 44, 12 (2016), 1700–1702.

[4] Janaína Gomide, Adriano Veloso, Wagner Meira Jr, Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. 2011. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proceedings of the 3rd international web science conference*. ACM, 3.

[5] Sahil Raj and Tanveer Kajla. 2015. Sentiment analysis of Swachh Bharat Abhiyan. *International Journal of Business Analytics and Intelligence* 3, 1 (2015), 32.

[6] Daniel Scanfeld, Vanessa Scanfeld, and Elaine L Larson. 2010. Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control* 38, 3 (2010), 182–188.

[7] Minesh P Shah, Benjamin A Lopman, Jacqueline E Tate, John Harris, Marcelino Esparza-Aguilar, Edgar Sanchez-Uribe, Vesta Richardson, Claudia A Steiner, and Umesh D Parashar. 2017. Use of Internet search data to monitor rotavirus vaccine impact in the United States, United Kingdom, and Mexico. *Journal of the Pediatric Infectious Diseases Society* (2017), pix004.

[8] Devendra K Tayal and Sumit K Yadav. 2017. Sentiment analysis on social campaign Swachh Bharat Abhiyan using unigram method. *AI & SOCIETY* 32, 4 (2017), 633–645.

[9] C Lee Ventola. 2014. Social media and health care professionals: benefits, risks, and best practices. *Pharmacy and Therapeutics* 39, 7 (2014), 491.