Chapter 8

Big Data Applications in Health Sciences and Epidemiology

Saumyadipta Pyne[*],[†],[1], Anile Kumar S. Vullikanti[‡], Madhav V. Marathe[§],[‖]

[*] *Bioinformatics, CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, University of Hyderabad Campus, Hyderabad, India*

[†] *Public Health Foundation of India, New Delhi, India*

[‡] *Computer Science and Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA*

[§] *Department of Computer Science, Virginia Tech, Blacksburg, Virginia, USA*

[‖] *Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA*

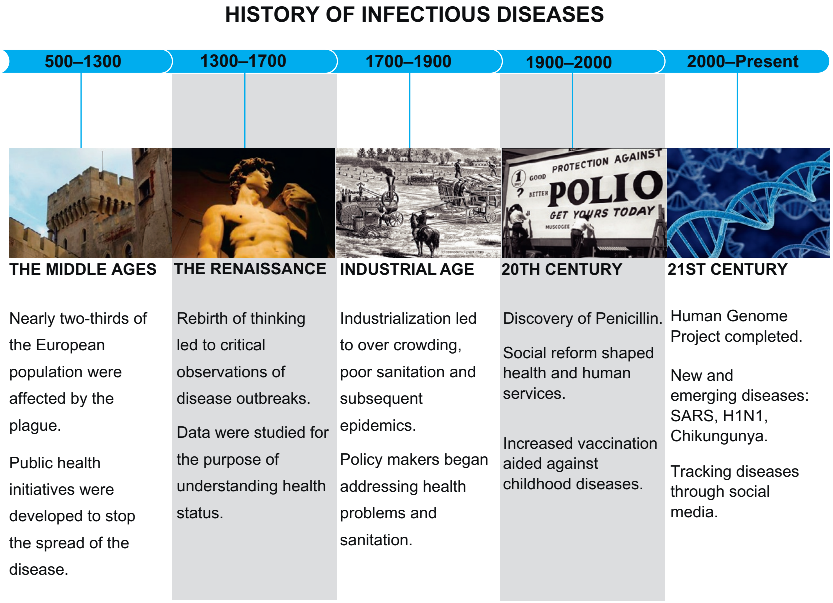
[1] *Corresponding author: e-mail: spyne@broadinstitute.org*

ABSTRACT

There is growing concern about our preparedness for controlling the spread of pandemics such as H1N1 Influenza. The dynamics of epidemic spread in large-scale populations are very complex. Further, human behavior, social contact networks, and pandemics are closely intertwined and evolve as the epidemic spread. Individuals' changing behaviors in response to public policies and their evolving perception of how an infectious disease outbreak is unfolding can dramatically alter normal social interactions. Effective planning and response strategies must take these complicated interactions into account. Mathematical models are key to understanding the spread of epidemics. In this chapter, we discuss a recent approach of diffusion in network models for studying the complex dynamics of epidemics in large-scale populations. Analyzing these models leads to very challenging computational problems. Further, using these models for forecasting epidemic spread and developing public health policies leads to issues that are characteristic of big data applications. The chapter describes the state of the art in computational and big data epidemiology.

Keywords: Mathematical epidemiology, Stochastic diffusion models, Public health policy planning, Epidemic surveillance, Epidemic forecasting, Agent-based simulations

Handbook of Statistics, Vol. 33. http://dx.doi.org/10.1016/B978-0-444-63492-4.00008-3



HISTORY OF INFECTIOUS DISEASES

500–1300	1300–1700	1700–1900	1900–2000	2000–Present
THE MIDDLE AGES	THE RENAISSANCE	INDUSTRIAL AGE	20TH CENTURY	21ST CENTURY
Nearly two-thirds of the European population were affected by the plague. Public health initiatives were developed to stop the spread of the disease.	Rebirth of thinking led to critical observations of disease outbreaks. Data were studied for the purpose of understanding health status.	Industrialization led to over crowding, poor sanitation and subsequent epidemics. Policy makers began addressing health problems and sanitation.	Discovery of Penicillin. Social reform shaped health and human services. Increased vaccination aided against childhood diseases.	Human Genome Project completed. New and emerging diseases: SARS, H1N1, Chikungunya. Tracking diseases through social media.

FIGURE 1 History of infectious diseases. Image courtesy: Maureen Lawrence-Kuether, Virginia Tech.

1 INTRODUCTION

Infectious diseases are the largest cause of human mortality worldwide—they account for more than 13 million deaths a year. Further, just six deadly infectious diseases—pneumonia, tuberculosis, diarrheal diseases, malaria, measles, and more recently HIV/AIDS—account for half of all premature deaths, killing mostly children and young adults. For example, malaria is said to be the primary cause of between 650,000 and 1.4 million deaths just in 2010. Key events related to the history of infectious diseases can be traced as far back as the middle ages (see Fig. 1).[1]

An epidemic is an occurrence in a community or region, of cases of an illness, specified health behavior or other health-related events in excess of what would be expected normally. The word epidemic is derived from two Greek works: *epi* (upon) and *demos* (people) meaning “upon people.” A pandemic is an epidemic that spans a large portion of the world, such as the H1N1 outbreak in 2009. In contrast, an endemic disease is one wherein new infections are constantly occurring in the population. Epidemics are thought to have influenced significant historical events including the plagues in Roman times and middle ages, the fall of the Han empire in the third century in China, and the defeat of the Aztecs in the 1500s, due to smallpox outbreak. The 1918 pandemic was responsible for more deaths than those due to World War I. The last 50 years have

[1] See a compilation of these events on wikipedia: http://en.wikipedia.org/wiki/List_of_epidemics.

seen epidemics caused HIV/AIDS, SARS, and influenza-like illnesses (see, e.g., Brauer et al., 2008).

Epidemiology is the study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the prevention and control of health problems (Last, 2001). By its very definition, the focus of epidemiology is on population-level issues as opposed to medical sciences, which is individual centric. Epidemiologists are primarily concerned with public health, which includes the design of studies, evaluation and interpretation of public health data, and the maintenance of data collection systems. They are interested in studying questions such as (1) What is the spatial extent of the outbreak? (2) How does the disease progress and how can we control it? (3) How did the disease originate and how does this compare to past outbreaks? Such questions have been asked repeatedly throughout human history. The 2014 Ebola outbreak in West Africa is an example of an epidemic that has caused thousands of deaths, along with significant social and economic impact on the affected countries.[2]

An important concern in epidemiology is to control an outbreak. With the advent of modern science, pharmaceutical measures have been widely used to control and prevent outbreaks. For example, vaccines have become a critical method of controlling, preventing, and possibly eradicating infectious diseases in host populations. Despite their success, nonpharmaceutical methods, such as quarantining and social distancing, continue to play a central role in controlling infectious disease outbreaks; they are especially important during an outbreak caused by an emerging pathogen. In such a case, pharmaceutical methods are usually not available.

Public health authorities have made significant strides in reducing the burden of infectious diseases. Nevertheless, infectious diseases continue to be an important source of concern. A number of global trends further amplify these concerns: (i) increased urbanization, (ii) increased global travel, (iii) denser urban regions, (iv) climate change, and (v) increased older and immunocompromised population. Many of the changes that we see around us are, to a large extent, anthropogenic and are happening at a scale wider and faster than ever before in human history. Further, new pathogens are emerging regularly, which raises the importance of societies' need to understand and be prepared to systematically address the challenge of emerging infectious diseases at different levels.

1.1 Mathematical and Big Data Computational Epidemiology

Mathematical models play an important role in epidemiology. Their importance is further highlighted by controlled physical experiments used to understand scientific phenomena being much harder and often impossible in epidemiology due to ethical and practical reasons. The first mathematical model in epidemiology

[2] See http://www.nytimes.com/interactive/2014/07/31/world/africa/ebola-virus-outbreak-qa.html.

is credited to Daniel Bernoulli in 1760 (Brauer et al., 2008). Using mathematical techniques, Bernoulli established that variolation[3] could help increase the life expectancy in the French population. Another systematic and data-driven investigation of the spread of epidemics was by John Snow, a British physician, who analyzed a cholera outbreak in London in 1854 and attributed it to a source of contaminated water (Brauer et al., 2008).

The early 1900s saw seminal advances in mathematical epidemiology. In 1911, Sir Ronald Ross discovered the malaria vector and transmission due to a species of mosquitoes and later developed a spatial model for the spread of the disease. One of the most significant results from his model was that the spread of malaria could be controlled by reducing the population of malaria below a "threshold"—this is the first instance of the concept of an epidemic threshold. Kermack and McKendrik extended this to develop the first general epidemic model, popularly known as the SIR model, involving ordinary differential equations (ODEs) based on a mass–action model (Brauer et al., 2008; Newman, 2003); we discuss this later in Section 2.

Big Data Computational Epidemiology is an emerging interdisciplinary area that uses computational models and big data for understanding and controlling the spatiotemporal diffusion of disease through populations. Figure 2 illustrates the four key components of variety, velocity, volume, and veracity in epidemiology. The role of computation and big data in epidemiological science has been progressively increasing over the last 20 years. There are several reasons for this change. First, mathematical models have become increasingly complex; it is virtually impossible to solve these models without the use of computers. Second, there is an increasing acceptance of networked models—these models represent the underlying population as a complex interaction network and study the spread of diseases over these interaction networks. These network models are more natural than the traditional compartmental models, but this comes at an increased computational and data cost. For example, the network models we discuss later in Section 6 integrate over 34 public and commercial data sets, leading to over 100 GB of input data for constructing a population model for the USA, with over 300 million people and 100 million locations. In turn, analysis of such data sets requires very powerful computing resources, and novel high-performance computing approaches (Barrett et al., 2008; Bisset et al., 2009a). Often, many parameters of the models are not known, especially in the early days of an epidemic (e.g., the transmission probability), necessitating parametric studies. Additionally, most policy questions involve comparison and evaluation of different interventions to control epidemics, such as vaccinations and school closures. Individual behavior can have significant impact on the

[3] Interestingly, it appears that the idea of *variolation* to control smallpox was known to Indians and Chinese as early as the eighth century AD. This method involved exposing people to material from smallpox scabs from infected individuals; see http://en.wikipedia.org/wiki/Inoculation for more details.

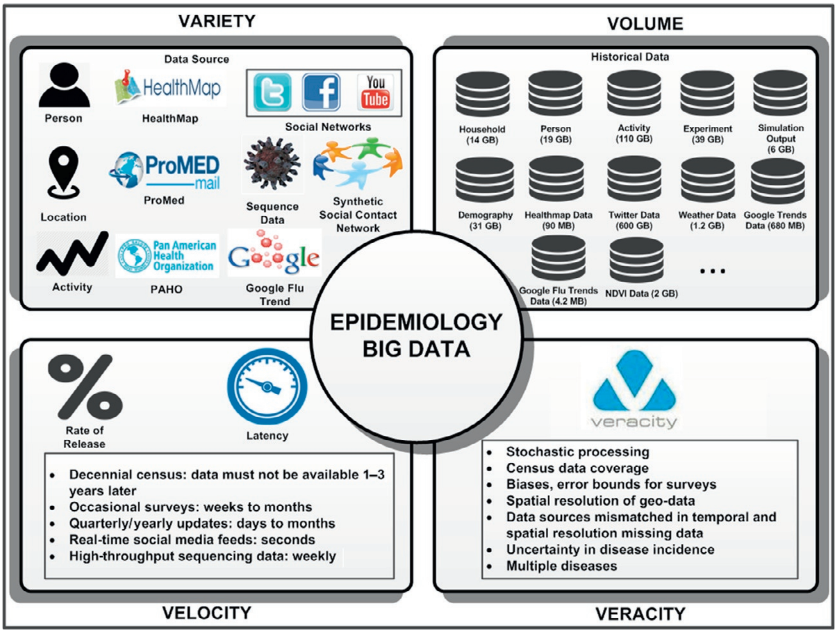
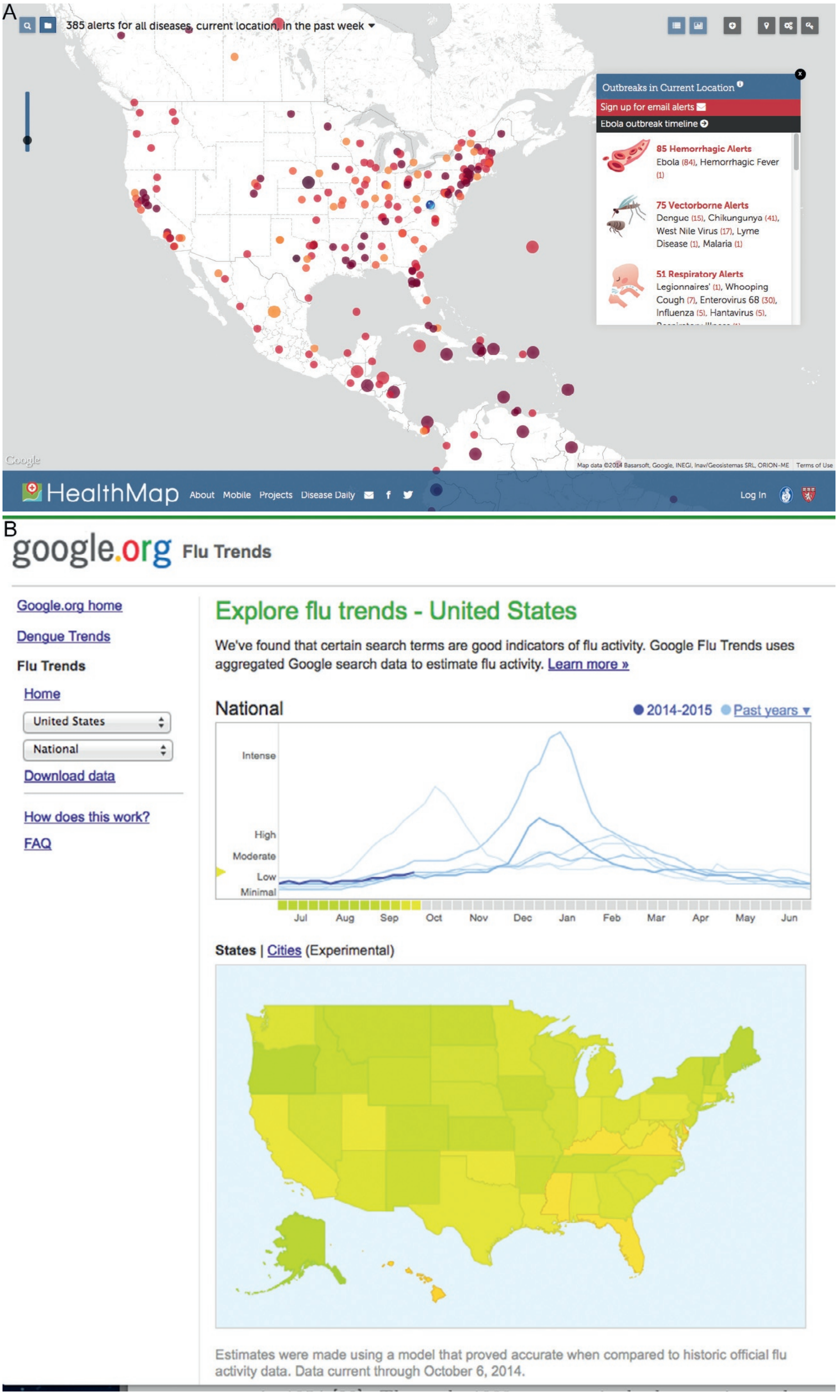

VARIETY
Data Source
Person
HealthMap
HealthMap
You Tube
Social Networks
Location
ProMED mail
ProMed
Sequence Data
Synthetic Social Contact Network
Activity
Pan American Health Organization
PAHO
Google
Google Flu Trend
VOLUME
Historical Data
Household (14 GB)
Person (19 GB)
Activity (110 GB)
Experiment (39 GB)
Simulation Output (6 GB)
Demography (31 GB)
Healthmap Data (90 MB)
Twitter Data (600 GB)
Weather Data (1.2 GB)
Google Trends Data (680 MB)
Google Flu Trends Data (4.2 MB)
NDVI Data (2 GB)
...
EPIDEMIOLOGY BIG DATA
Rate of Release
Latency
Decennial census: data must not be available 1–3 years later
Occasional surveys: weeks to months
Quarterly/yearly updates: days to months
Real-time social media feeds: seconds
High-throughput sequencing data: weekly
veracity
Stochastic processing
Census data coverage
Biases, error bounds for surveys
Spatial resolution of geo-data
Data sources mismatched in temporal and spatial resolution missing data
Uncertainty in disease incidence
Multiple diseases
VELOCITY
VERACITY

FIGURE 2 Big data issues in epidemiology: variety, volume, velocity, and veracity. Image courtesy: S.M. Shamimul Hasan, Virginia Tech.

disease dynamics (Funk et al., 2010; Lipsitch et al., 2009) and further increase the number of instances that need to be done. Third, epidemiologists are collecting increased amounts of data from the field. New methods are being developed for disease surveillance and detection. Although extremely valuable, the data need to be managed at all levels. Computational methods for data management, including methods to collect, store, clean, organize, search, fuse, and analyze data, are all important. Finally, with the advent of the world wide web, there is a growing demand for developing web-based tools that can be accessed by epidemiologists in a pervasive manner. An example of this is the HealthMap tool, which collects data about disease incidence all over the world, through news and social media data (Fig. 3a). Another tool is Google's FluTrends, which infers flu incidence from search queries, e.g., related to flu medication, news, etc. (Fig. 3b). Overall, big data play a large and important role in epidemiology (Salathé et al., 2012).

1.2 Organization and Outline

This chapter describes the basic concepts in computational epidemiology, with a focus on networked epidemiology. We discuss different kinds of mathematical models for epidemiology in Section 2. Some of the common computational



Outbreaks in Current Location

Sign up for email alerts

Ebola outbreak timeline

85 Hemorrhagic Alerts
Ebola (84), Hemorrhagic Fever (1)

75 Vectorborne Alerts
Dengue (15), Chikungunya (41), West Nile Virus (17), Lyme Disease (1), Malaria (1)

51 Respiratory Alerts
Legionnaires' (1), Whooping Cough (7), Enterovirus 68 (30), Influenza (5), Hantavirus (5),

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through October 6, 2014.

FIGURE 3 Screenshot from (a) HealthMap (http://www.healthmap.org/en/) and (b) Google Flu Trends (http://www.google.org/flutrends/), taken on October 7, 2014.

problems of analysis and inference are discussed in Sections 3 and 4, respectively. Section 5 discusses concepts in molecular epidemiology, disease surveillance, and phylodynamics and outlines important big data challenges in these fields. Section 6 discusses a computational framework for networked epidemiology. This includes a synthetic information environment for organizing and reasoning about diverse data sets, high-performance computing-based modeling environments, and pervasive web-based decision support tools.

2 MATHEMATICAL FRAMEWORK FOR EPIDEMIOLOGY

We start with a very commonly used model, popularly known as the SIR model, involving ODEs based on a mass–action assumption; we refer to Brauer et al. (2008) and Newman (2003) for more details. A population of size N is divided into three states: susceptible (S), infective (I), and removed or recovered (R) (Fig. 4). The following process describes the system dynamics: each infected person can infect any susceptible person (independently) with probability β and can recover with probability γ. Let $S(t)$, $I(t)$, and $R(t)$ denote the number of people who are susceptible, infected, and recovered states at time t, respectively. Let $s(t) = S(t)/N$, $i(t) = I(t)/N$, and $r(t) = R(t)/N$; then, $s(t) + i(t) + r(t) = 1$. Under the "complete mixing" assumption that each individual is in contact with everyone in the population, the dynamics of the SIR model can be described by a system of differential equations (Fig. 4b).

One of the classic results in the SIR model is that there is an epidemic which infects a large fraction of the population, if and only if $R_0 = \beta/\gamma > 1$; we refer the reader to Dimitrov and Meyers (2010) for a derivation of this result. R_0 is known as the "reproductive number" and is key in characterizing an epidemic. At the start of an epidemic, much of the public health effort is focused on estimating R_0 from observed infections (Lipsitch et al., 2003). Further, public health decision making for controlling an epidemic is often translated to reducing R_0—this can be done in many ways, such as by reducing β through better protective measures, such as washing and face masks, in the case of flu (Dimitrov and Meyers, 2010). While the notion of R_0 has given a lot of insights into the spread of epidemics, it has several limitations, especially in practice. For instance, during the SARS outbreak in 2003, R_0 was estimated to be in the range $[2.2, 3.6]$. However, despite its spread over many countries, it

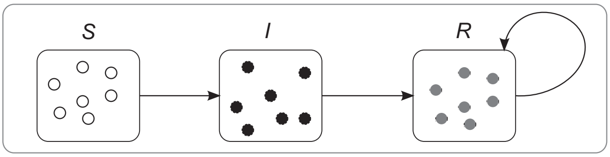


FIGURE 4 The SIR model. (a) The entire population is split into compartments corresponding to the S, I, and R states. Any two individuals in the population can come in contact. (b) Coupled differential equations for the SIR model.

did not cause a very large outbreak for multiple reasons. First, the R_0 estimates were based on infections in crowded hospital wards, where a complete mixing assumption is reasonable. But this does not hold in other settings. Second, as the epidemic spread, people altered their behavior and started mixing less, which is not taken into account in the static definition of R_0.

There has been a lot of work in developing more complex compartmental models in order to address some of these issues—compartments represent subpopulations with very similar characteristics and are well mixed within themselves so that the dynamics can be expressed in terms of differential equations over these groups. See Newman (2003) and Brauer et al. (2008) and the references therein for more details. Overall, compartmental models have been immensely successful over the last 100 years and have become a workhorse of mathematical epidemiology. They are relatively easy to extend and quick to build and can either be solved analytically or numerically quite efficiently, building on the well-developed theory of ODEs.

In contrast to compartmental models, *networked epidemiology* considers epidemic spread on a specific undirected contact network $G(V, E)$ on a population V—each edge $e = (u, v) \in E$ implies that individuals (also referred to as nodes) $u, v \in V$ come into contact.[4] Let $N(v)$ denote the set of neighbors of v. For instance, in the graph in Fig. 5, we have $V = \{a, b, c, d\}$ and $E = \{(a, b), (a, c), (b, d), (cd)\}$. Node a has b and c as neighbors, so $N(a) = \{b, c\}$. The SIR model on the graph G is a dynamical process in which each node is in one of S, I, or R states. Infection can potentially spread from u to v along edge $e = (u, v)$ with a probability of $\beta(e, t)$ at time instant t after u becomes infected, conditional on node v remaining uninfected until time t—this is a

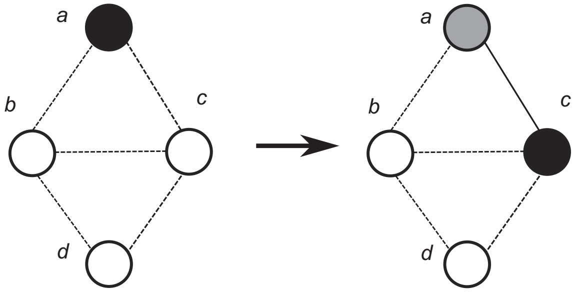


FIGURE 5 The SIR process on a graph. The contact graph $G = (V, E)$ is defined on a population $V = \{a, b, c, d\}$. The node colors white, black, and gray represent the Susceptible, Infected, and Recovered states, respectively. Initially, only node a is infected, and all other nodes are susceptible. A possible outcome at $t = 1$ is shown, in which node c becomes infected, while node a recovers. Node a tries to independently infect both its neighbors b and c, but only node c gets infected—this is indicated by the solid edge (a, c). The probability of getting this outcome is $(1 - \beta((a, b), 1))\beta((a, c), 1)$.

[4] Note that though edge e is represented as a tuple (u, v), it actually denotes the set $\{u, v\}$, as is common in graph theory.

discrete version of the rate of infection for the ODE model discussed earlier. We let $I(t)$ denote the set of nodes that become infected at time t. The (random) subset of edges on which the infections spread represents a disease outcome and is referred to as a *dendrogram*. This dynamical system starts with a configuration in which there are one or more nodes in state I and reaches a fixed point in which all nodes are in states S or R. Figure 5 shows an example of the SIR model on a network.

The time series $(|I(t)|, t = 0, 1, \ldots)$ is referred to as an *epidemic curve* corresponding to a stochastic outcome; this is a very commonly used quantity related to an epidemic. The *total number of infections* for an outcome is given by $\sum_t |I(t)|$. The *peak value* is $\max_t |I(t)|$, while the *time to peak* is the largest time t that maximizes $|I(t)|$. These quantities are illustrated in Fig. 6. Also observe that the epicurve is not very smooth; this is in contrast with the epicurve in an ODE model. A popular variant of the SIR model is the SIS model, in which an infected node switches to state S after the infectious duration. Some of the notation used in the paper is summarized in Table 1.

Such a network model captures the interplay between the three components of computational epidemiology: (i) individual behaviors of agents; (ii) unstructured, heterogeneous multiscale networks; and (iii) the dynamical processes on these networks. It is based on the hypothesis that a better understanding of the characteristics of the underlying network and individual behavioral adaptation can give better insights into contagion dynamics and response strategies. Although computationally expensive and data intensive, network-based epidemiology alters the types of questions that can be posed, providing qualitatively different insights into disease dynamics and public health policies. It also allows policy makers to formulate and investigate potentially novel- and context-specific interventions. Such epidemic models are instances of a more general dynamical systems framework, called *graph dynamical systems* (GDSs) (Mortveit and Reidys, 2007). Therefore, the theory of GDS can be used as a

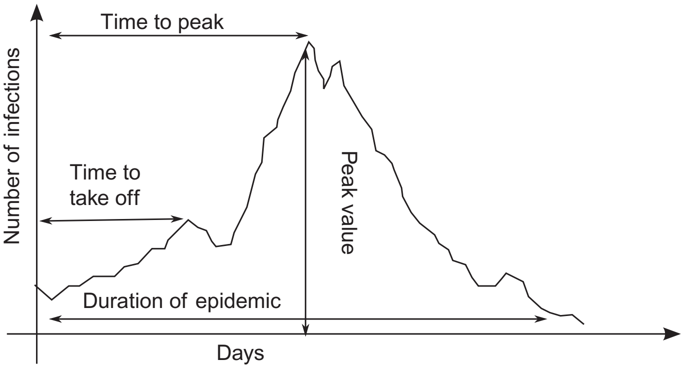


FIGURE 6 Different quantities associated with the dynamics of epidemics. See discussion in Section 2 for definitions of these quantities.

TABLE 1 Summary of Some of the Notations Used in the Chapter (See Marathe and Vullikanti, 2013 for More Discussion)

$G(V, E)$	Social contact network on a population V, with each edge $e = (u, v) \in E$ representing a contact
$N(v)$	Set of neighbors of node $v \in V$ in graph G
$\beta(e, t)$ for $e = (u, v)$	Transmission probability: the probability that the infection spreads from u to v per time unit of contact, if u is infected and v is not
$I(t)$	The set of nodes infected at time t
$(\lvert I(t)\rvert, t = 0, 1, \ldots)$	Epidemic curve
$\max_t \lvert I(t)\rvert$	Peak value of an epidemic
$\text{argmax}_t \lvert I(t)\rvert$	Time to peak

foundation for computational epidemiology; this leads to four basic classes of problems, which are motivated from the analysis of dynamical systems:

1. *Analysis problems.* This includes characterizing the outbreak size, duration of the epidemic, and other properties of epidemics. There has been a lot of work on analytical results in terms of network properties, e.g., Marathe and Vullikanti (2013), Easley and Kleinberg (2010), Newman (2003), Pastor-Satorras and Vespignani (2001), Alon et al. (2004), Chung and Lu (2002), Ganesh et al. (2005), and Wang et al. (2003), and on computational tools for such analysis, e.g., Perumalla and Seal (2011), Carley et al. (2006), Parker and Epstein (2012), Eubank (2002), Ferguson et al. (2006), Longini et al. (2005), Germann et al. (2006), and Chao et al. (2010). We discuss these problems in more detail in Section 3.
2. *Inference problems.* As discussed earlier, the networked SIR model is determined by the network, initial conditions, and epidemic model. In general, we may have partial information about some of these components; e.g., edges of the graph might not be completely known, or parameters of disease spread are unknown. Some of the results on such inference problems include Nishiura (2011a), Neil et al. (2005), Nsoesie et al. (2011), Shah and Zaman (2010), Netrapalli and Sanghavi (2012), and Gomez-Rodriguez et al. (2010). These are discussed in more detail in Section 4.
3. *Optimization problems.* This includes problems of controlling the spread of epidemics, e.g., by vaccination or quarantining, correspond to making changes in the node functions or removing edges so that the system converges to configurations with few infections, e.g., Borgs et al. (2010), Aspnes et al. (2006), and Eubank et al. (2004a). These are very challenging stochastic optimization problems, and beyond the scope of this chapter. We refer the reader to Marathe and Vullikanti (2013) for some pointers to this area.

4. *Forecasting and situational assessment problems.* These include problems of determining quantities, such as the number of infections over time, or the peak, and identifying the people who might be infected, given partial information of the outbreak till some time. This is a very active area of research (see, e.g., Chakraborty et al., 2014a; Ginsberg et al., 2009). See Section 5.2 for more details.

3 DYNAMICS AND ANALYSIS PROBLEMS

The structure of the underlying contact graph G has a significant impact on the disease dynamics (Newman, 2003). A fundamental question is to characterize the conditions under which there is a "large" outbreak, which is defined as one which infects $\Theta(n)$ individuals, where $n = |V|$. We mention a few of the main results of this kind for the SIR and SIS models. These are somewhat involved to derive here, and we give references for more details. A common approach is to try to characterize the dynamics in terms of the degree distribution of the graph. The simplest case is when G is a complete graph, with a uniform probability β of the disease spread from u to v, and $\tau(u) = 1$. The classical result by Erdős-Renyi (Marathe and Vullikanti, 2013; Newman, 2003) implies that there is a large outbreak with $\Theta(n)$ infections, if and only if $\beta > 1/n$. The main technique is a branching process analysis. We note that this result for the complete graph is a discrete analogue of the characterization in terms of R_0, with $n\beta$, the expected number of infections caused by a node, being the equivalent of R_0. The existence of such a threshold for disease spread has been shown to exist in other well-structured graph classes, such as lattices and random regular graphs (Newman, 2003). It is much more challenging to prove such statements in heterogeneous graphs with less symmetry. Pastor-Satorras and Vespignani (2001) use techniques from statistical mechanics to show that in scale-free network models, under mean-field assumptions the threshold for epidemics propagation is 0, i.e., no matter how small the probability of infection is, there would be a large outbreak. Two settings for which rigorous results are known, without using any mean-field assumptions, are (i) the Chung-Lu model (Chung and Lu, 2002), which is a random graph model in which the probability of an edge (i, j) is proportional to $w_i \cdot w_j$, for a given weight sequence $\mathbf{w}$; and (ii) classes of expander graphs (Alon et al., 2004).

4 INFERENCE PROBLEMS

In general, not all elements of the network model are known. For instance, the network might be partially known. Similarly, the transmission probability and the initial configuration, such as the source of the infection might not be known. Partial information in the form of observed infections might be available through public health surveillance. At the onset of an outbreak (e.g., in every flu season), recurring problems for public policy planners include determining where the epidemic might have started, characteristics of the specific disease strain, and if there will be a large epidemic or pandemic. These are significant data and computational challenges, where statistical and machine learning approaches have an important role.

General abstractions of this problem are to determine the transmission probability or other disease characteristics, or the probability that the source of the epidemic is a node v, conditioned on observed surveillance data and some assumptions about the prior distribution, or to find a maximum likelihood estimator for them. A related class of problems is to infer the underlying contact network properties, based on such observations. These are very challenging problems because (i) surveillance data on who is infected are limited and noisy—only a fraction of the infected people are detected; (ii) the underlying contact graph is only partially known and is dynamic. These problems are the "inverse" of the "forward" problems of simulating the spread of an epidemic in a social network.

There has been a lot of work related to these problems, especially in the statistics and data mining literature (Neil et al., 2005; Nishiura, 2011a; Nsoesie et al., 2011). An active area of research is based on Bayesian approaches, e.g., using spatial scan statistics for detecting anomalous outbreaks (Neil et al., 2005). These techniques rely on detecting spatial clusters from syndromic surveillance data, without using any specific properties about the epidemic process. In contrast, a model-based approach relies on the characteristics of the SEIR process, and Nsoesie et al. use a combination of simulations and supervised classification techniques to predict epidemic curves and infer underlying disease parameters for an ongoing outbreak (Nsoesie et al., 2011).

One of the problems that have been studied is inferring the source of an infection, given the network and observed infections. This is generally formalized as a likelihood maximization problem. Shah and Zaman (2010) develop an efficient maximum likelihood approach for determining the source of an outbreak in trees for an SI model of disease and then extend it to general graphs by assuming spreading in a Breadth First Search (BFS) order. They also develop a heuristic, called rumor centrality, which allows faster estimation of the source probability. The related problem of inferring the contact network based on the disease spread information has been studied by Netrapalli and Sanghavi (2012) and Gomez-Rodriguez et al. (2010), who present maximum likelihood estimators for a minimal network on which a given set of observed infections can occur.

5 DISEASE SURVEILLANCE, MOLECULAR EPIDEMIOLOGY, AND PATHOGEN PHYLODYNAMICS

The emergence of new pathogens has led to new, critical challenges that are faced by the hitherto restricted interfaces of humans, animals, and the environment (Patz et al., 2004). Many of such interfaces around the world have been identified as hot spots of emerging zoonotic infectious diseases, in which an array of ever-increasing human activities such as farming, mining, logging, and hunting are being conducted against the backdrop of significant wildlife biodiversity, of course, with concomitant microbial diversity (Morse et al., 2012). Perhaps the regions in which environmental changes have been occurring most rapidly, and where people and wildlife live in close proximity, are located mostly in the developing world, and which, ironically, have limited resources and capacity to monitor and respond to zoonotic outbreaks.

Morse et al. (2012) described a three-stage model for typical dynamics of transmission of zoonotic diseases from a localized hot spot to a global pandemic. In preemergence stage 1, a pathogen is localized in its natural animal reservoir but could be driven by certain major ecological or environmental disturbances, which increases the likelihood of its coming to contact with, and spilling over to, new species including livestock or humans that are in increasingly close contact. For instance, the recent outbreak of Ebola has been brought back to focus the issue of "growing human activity and deforestation in previously untouched forests."[5] The 2002 SARS outbreak in China was traced to hunting and trading of bats that are natural reservoirs of SARS-like corona viruses (Li et al., 2005).

In stage 2, while the disease spills over to humans, it remains localized within a geographical region with outcomes varying between a few cases to a moderate or even large outbreak, but possibly with limited interpersonal transmission, as, for instance, in the case of the recent Ebola outbreak. A large-scale pandemic occurs in stage 3 and is quite rare. It is usually attributed to global travel and trade and characterized by sustained interpersonal transmission and international spread of reservoir hosts or vectors. To address the problem of zoonosis, clearly the most desirable objective is to tackle and limit the transmission of a zoonotic pathogen, if possible, in stage 1. Unfortunately, the global trend in emerging infectious diseases has seen increasing incidence since 1940, despite controlling for effects of higher reporting (Jones et al., 2008). It was found that 60% of known human infectious diseases, and more than 75% of emerging infections, have zoonotic origin (King et al., 2006; Taylor et al., 2001). In fact, 71.8% of zoonotic pathogens originated in wildlife species. Approximately 80% of the identified reservoirs are mammalian, and almost half of the observed 1000 pathogens noted in livestock and pets are zoonotic. Indeed, almost 80% of the viruses and 50% of the bacteria that infect humans are of zoonotic origin. A conservative estimate puts the number of viruses present in vertebrate species at 1 million, which suggests that more than 99.9% of the viruses are currently unknown (Lipkin, 2009). While such staggering numbers are a matter of great concern, it also presents a perspective to the researchers so that they may be prepared to approach the problem in a comprehensive and systematic manner. Data on such pathogens and their evolution help in disease surveillance and

[5] http://www.washingtonpost.com/news/morning-mix/wp/2014/07/08/how-deforestation-and-human-activity-could-be-to-blame-for-the-ebola-pandemic/.

forecasting and lead to significant big data challenges, as illustrated in Fig. 2. We briefly summarize these aspects below.

5.1 Disease Surveillance

Recently, the concept of "One Health" (Coker et al., 2011) has received much attention led by the combined initiatives of FAO and the UN, WHO, and the World Organization for Animal Health. Importantly, the need for a concerted, interdisciplinary approach has been felt in public health research, bringing together teams of epidemiologists, clinicians, veterinarians, microbiologists, wildlife biologists, ecologists, data analysts, and public health officials. Apparently, a global strategy seems to be emerging for monitoring and biosurveillance, partly driven by concerns of bioterrorism. However, the coverage or efficiency of public health surveillance systems worldwide is still far from uniform.

Notable efforts in this direction are made with the help of social media (Lipkin, 2013). ProMED (Program for Monitoring Emerging Infectious Diseases) was created in 1994. In a bottom-up manner, a network of reader at the grassroots submits entries that are then curated by an expert panel that sends ProMED-mail with commentary in five languages to a listserv exceeding 60,000 subscribers in 185 different countries. A private subscription-based service GPHIN (Global Public Health Intelligence Network) receives information on global outbreaks by scanning worldwide news in nine languages. In 2001, GPHIN joined the WHO Global Outbreak Alert Response Network (GOARN), which included additional verification services to its reporting. HealthMap.org, launched in 2006, is a user-friendly map-based online display system for real-time crowd-sourced submission of reports of worldwide georeferenced observations from news media, either through its web site or through a number of smart-phone-based apps (e.g., Outbreaks Near Me). Figure 3 shows screenshots of HealthMap and Google FluTrends, which are two tools for disease surveillance.

Typically, automated surveillance mechanisms would generate data with the big data characteristics of high volume, velocity, and variety. Researchers in big data and Computational Statistics must design robust models and algorithms to address the analytical needs of working with such data. A variety of methodological advances have been made to analyze data from social networks, sensor networks, data streams, GIS data, and associated platforms now routinely used for monitoring and surveillance (Cooper et al., 2006). Computational frameworks for biosecurity involve clustering, classification, prediction, outlier analysis, univariate and multivariate stream analysis, data fusion, social network analysis, text and web data mining, spatial and spatiotemporal modeling, and visual analytics, among others.

Many public health researchers and policy experts are currently involved in designing programs and strategies to be prepared for emerging (and reemerging) diseases. Systematic identification of new hot spots; modeling and mapping of ecological niches; characterization of human–animal interfaces by exposure types and frequency; creation of host–pathogen databases; identification of key taxonomic groups; and phylodynamic analysis of host, epidemiological, and molecular data all lead to a pool of information that could be useful for prediction and prevention of zoonotic outbreaks. As an example of this approach, the PREDICT project of the Emerging Pandemic Threats (EPT) program launched by US Agency for International Development (USAID) in 2009 aims to facilitate predictive modeling for the identification of the most likely regions, hosts, and human–animal interfaces for forthcoming emergence of zoonoses (Morse et al., 2012).

As described by Morse et al. (2012), PREDICT aims to collect timely and reliable data based on Internet surveillance of reports of unusual health events occurring in countries with hot spots. Further, it conducts analysis to test if the corresponding pathogen is likely to emerge and spread in the social systems that exist in those hot spots. PREDICT combines risk modeling with targeted wildlife field sampling for selected locations, interfaces, and host taxa. In this effort, it is aided by interdisciplinary experts, computerized data collection and analysis, and active partnership with local and national governments. The program currently collaborates with 20 African, Asianm and South American countries, and in just a few years, it has detected hundreds of novel viruses in the hundreds of thousands of samples collected from tens of thousands of animals from these locations (Morse et al., 2012).

While programs such as PREDICT can demonstrate the benefits of local capacity-building efforts (Schwind et al., 2014) as part of international efforts to counter zoonotic threats, it also underscores the critical need for statistical and computational capability to work with massive data sets of high volume, velocity, and variety. To create sophisticated models for forecasting, the researchers must consider high-dimensional data on a large number of parameters: socioeconomic drivers (e.g., population growth and density, mixing patterns, migration, trade, agricultural practices, sanitation, age, diet, vaccination, drug and antibiotic use, cultural norms, occupational exposures, nutritional and immunological status), wildlife diversity, contact frequency, relatedness of host species, relatedness of microbial species present in host, evolvability of pathogens, host–pathogen coevolution, and several general factors such as capacities for reporting and response on the ground, geographical, and ecological conditions, and so on. An array of tools designed for predictive analytics, ranging from neural networks and tree models and regression models based on the observed parameters, to complex representations such as Bayesian network models that can capture their underlying dependence structure, are gaining in popularity (Buczak et al., 2013; Cooper et al., 2006).

Serological surveillance is routinely conducted worldwide in order to determine the prevalence of a disease in a population. From a statistical perspective, cross-sectional or longitudinal serological surveys provide useful data in terms of levels of antibodies in serological samples owing to past infections. Thus, parametric or nonparametric models are generally used to separate the susceptible and the infected subpopulations based on seroprevalence data, within specific age groups or otherwise, allowing inference on such parameters as the prevalence and the rate of infection, as described in Hens et al. (2012).

5.2 Forecasting

An important emerging topic in big data computational epidemiology is epidemic forecasting. The basic idea here is to combine big data obtained from nontraditional sources, including social media, wikipedia, electronic health records, crowd-sourcing as well as surveillance systems with causal and statistical models to *nowcast* and *forecast* prevalence of diseases in the host population. Just like weather systems and markets, epidemics are complex systems; as a result forecasting the future incidence rates is challenging. See Nsoesie et al. (2013), Nishiura (2011a), Ohkusa et al. (2011), Hall et al. (2007), and Tizzoni et al. (2012) for a recent survey and early work on this topic. Shaman et al. have used Bayesian ensemble methods to develop surprisingly high-quality forecasts for flu prevalence in US regions (Shaman and Karspeck, 2012; Shaman et al., 2010a,b). An important contribution of their work was to elucidate the role of weather (humidity) on the incidence rates. In a recent paper (Chakraborty et al., 2014b), as a part of the IARPA-funded EMBERS project, we analyzed the generation of robust quantitative predictions about temporal trends of flu activity, using several surrogate data sources for 15 Latin American countries. We also recently participated in the CDC flu challenge and used our models to develop real-time flu forecasts for the nine US regions. An important consideration in both cases was to produce predictions in real time before the event and compare these forecasts retrospectively using curated data sets from CDC and PAHO. See https://www.federalregister.gov/articles/2013/11/25/2013-28198/announcement-of-requirements-and-registration-for-the-predict-the-influenza-season-challenge. Recently, other agencies have also put out similar competitions, e.g., a similar competition for chikungunya was initiated by DARPA (see https://www.innocentive.com/ar/challenge/9933617). The challenges have helped focus renewed interest on developing innovative techniques for infectious disease forecasting.

5.3 Molecular Epidemiology

In the “postgenomic” era, a significant gain in the predictive value of such modeling comes from incorporating molecular data, which are used along with the traditionally available clinical, pathological, and epidemiological data. Classically, Koch’s postulates, formulated by Koch and Loeffler in 1884, have served as the gold standard criteria to establish a causative relationship between a microbe and a disease. They require that the agent must be present in all cases of the disease, it must be specific, and it must be isolated and cultivated as culture, which, if inoculated in a healthy host is, must be sufficient to reproduce the original disease. With the advent of molecular methods, these criteria were modified to allow sequence-based identification of microbial pathogens by Fredricks and Relman. Recently, "metagenomic Koch's postulates" were proposed such that metagenomic markers (individual sequence reads, assembled contigs, genes, or genomes) could be used to detect viruses, most of which are not easy to culture (Mokili et al., 2012). An integrative approach to establish causation was described by Lipkin (2013) in terms of possible, probable, and confirmed causal relationships.

Systematic identification of the causative agent can play a key role in epidemiology; it can help not only with estimation of the basic reproductive number (R_0) but also with inference of the probable routes and dynamics of transmission, and even indicate possible intervention strategies. Traditional approach to pathogen discovery has involved an established array of techniques ranging from *in vitro* and *in vivo* culturing to platforms such as immunohistochemistry, electron microscopy, and serology (Morse et al., 2012). In the 1990s, the advent of high-throughput "omic platforms revolutionized rapid, cost-effective molecular data generation, enabled by important methodological advances in statistics and bioinformatics, paved the way for unbiased, data-driven discoveries". Multiplex PCR arrays, mass spectrometry, microarrays, and high-throughput sequencing are now routinely used for microbial analysis. As costs (money, time need for analysis) of such analyses continue to fall, while their efficiency (multiplexing, throughput) continues to rise, routine screening of samples for prospective pathogen discovery could soon become a viable strategy to preempt the emergence of zoonotic disease in a systematic manner.

High density of primer sequences and the increasing robustness of the technology allow nucleotide-based microarrays (or "chips") to probe for a large number of potential viruses very cheaply and rapidly. In fact, with a flexibly designed set of probes, and the compactness of 8–10 million spots, the possibility of having a single chip that can test for all the known viruses, at least for narrowing down to the level of genus or family, for the price of just $20 or less, is not far-fetched (Vaidyanathan, 2011). The ViroChip was used to detect the presence of coronavirus family in the respiratory samples of SARS in less than 24 h and without the need for culturing the virus (Chen et al., 2011). Besides the pan-viral ViroChip, there are also the pan-microbial microarrays such as GreeneChip (Palacios et al., 2007) and LLMDA (Gardner et al., 2010). The trend is to use components that can make such platforms in the future even more specific in their detection capacity and smaller in size so that they could be more effective and easily deployed in the field. Parallel and distributed computing, on the other hand, could make search for agents and their biomarkers more robust and efficient during outbreaks, e.g., Fox et al. (2013).

The real game-changer in the field of novel microbe hunting has been, however, due to the advances in high-throughput sequencing (HTS) over the past decade (see reviews by Loman et al., 2012; Metzker, 2009). Currently, there are multiple HTS platforms available, each with their own strengths and weaknesses. The main reason that makes HTS so suitable for pathogen discovery is that unlike previous methods like the consensus PCR or microarrays which are dependent on the primer sequences, it allows unbiased identification of both known and unknown agents, including highly divergent and novel pathogens. A list of viral pathogens that were detected by HTS is given in Chiu (2013).

The first HTS platform to be thus used was the 454 Life Sciences pyrosequencer, which is still in use today as the Roche GS FLX Titanium and GS Junior systems. However, the more recent and currently the most widely used HTS systems are from Illumina, which include the massively parallel HiSeq 2000, and the faster but lower throughput MiSeq (Firth and Lipkin, 2013). The selection of an HTS system for pathogen discovery is determined by various factors. First, as the pathogen genomes, especially viral genomes, are relatively small compared to that of hosts, the amount of template required for input must not be prohibitive. In particular, in tissue or fecal samples, the ratio of viral to background material is often quite low. This leads to steps of careful steps toward enriching the agent (or host depletion) during sample preparation for HTS (Firth and Lipkin, 2013). The two key HTS parameters for pathogen discovery are the read length and the depth of coverage. While reads must be long enough (at least 100–300 bases) to allow unambiguous identification of the agent as distinctively as possible from the host or background sequences, the read depth (i.e., the number of sequence reads per run), on the other hand, must be reasonably high to allow easy assembly of the microbial genome and thus enable sensitive detection. Illumina platforms generate almost 10–1000 times more data per run with much smaller read lengths (HiSeq 600 Gb, 150 bases; MiSeq 3 Gb, 250 bases) as compared to GS FLX (0.7 Gb, up to 700 bases) but they also take significantly longer runtime (HiSeq 11 days, MiSeq 27 h, GS FLX 8 h) (Chiu, 2013; Firth and Lipkin, 2013). Clearly, running time and costs are critical factors, especially during an outbreak. Fortunately, the per-base cost of sequencing has fallen steadily, from $5000 per megabase in 2001 using classical dideoxy methods to $0.50 per megabase in 2012 using the Illumina platform (Lipkin, 2013). A virus that might have taken a week to sequence a decade ago (say, SARS-coronovirus in 2003) now requires at most a day to finish. Further, access to HTS labs is also gradually increasing worldwide.

Increasingly, it is not so much the constraints of data acquisition that are of concern to HTS labs across the globe as the challenge of efficient analysis of the large amounts of generated data. Traditionally, the computational pipelines for pathogen detection have involved the key steps of (a) computational subtraction of the host or background sequences using reference databases, and (b) BLAST homology search of the target agent against all known microbes. If BLASTn search by nucleotide homology does not work well for identification of distantly related or novel pathogens, then protein-level similarity with translated BLASTx query is used to explore further, especially for long reads exceeding 150 bases (Firth and Lipkin, 2013). More divergent relationships are explored with frameworks such as PHI-BLAST or HMMER (Finn et al., 2011). In the recent years, several analytical pipelines, such as PathSeq, CaPSID, and READSCAN, were developed specifically for HTS-based pathogen discovery, e.g., Kostic et al. (2011), Borozan et al. (2012), and Naeem et al., (2013). New pipelines, such as SURPI (Naccache et al., 2014), are getting more scalable and efficient by taking advantage of cloud-based technologies (in "fast mode," SURPI detects pathogens analyzing data with 7500 million reads in 11 min to 5 h).

As only less than 1% of the world's microbial diversity has been cultivated in the lab, an entire vista has been opened up by the recent advances in the field of metagenomics, which is a culture-independent approach for "functional and sequence-based analysis of collective microbial genomes contained in environmental samples" (Riesenfeld, 2004). Viral metagenomics is getting increasingly popular for a wide range of investigations which include not only environmental research areas, such as marine ecology or agricultural biotechnology, but also diagnostics of human diseases (Barzon et al., 2011). While it was originally based on cloning methods for analyzing double-stranded DNA genomes, the use of HTS now allows metagenomic analysis of all kinds of genomes, including that of single-stranded DNA and RNA (Mokili et al., 2012). Typically, a pipeline for metagenomic analysis consists of a sample preparation step, which helps in removal of nonviral nucleic acids, followed by HTS, and finally, bioinformatics analysis whereby the genomic sequences of individual microbes present in the collection are distinguished.

While the presence of individual microbes is detected post hoc from a mixture thereof in metagenomic analysis, an alternative and interesting approach is to first single out the agents and then amplify the nucleic acids and sequence them individually. Single-cell analysis (SCA), followed by metagenomics using HTS, can significantly improve the sensitivity of microbial detection (Blainey, 2013). In SCA, individual cells are first isolated from the sample, captured separately (say, in microfluidic chips), and then sent for metagenomic or metatranscriptomic analysis. This obviates the need for post hoc computational separation of the microbial genomes. While multiple projects in single-cell genomics have been conducted with bacterial species, recently single viral genomics (SVG) was conducted by Allen et al. (2011). Using platforms such as flow cytometry, single virions were selectively isolated before sequencing was carried out. Interestingly, by incorporating an intermediate reverse-transcription step prior to HTS, the SVG approach could be particularly useful in the genomic analysis of viral heterogeneity, especially for RNA viruses, which have much higher mutation rates during replication (and absence of error-correction), and thus form viral quasispecies (Holmes, 2009). With their high mutation and replication rates, and under selective pressures from the use of drugs and vaccines, RNA viruses can quickly evolve to use new cell receptors or routes of transmission in unexpected host species (Firth and Lipkin, 2013).

5.4 Pathogen Phylodynamics

A systematic understanding gained through sustained and continuous investigation of viral genetic diversity should be a key aspect of advanced biosurveillance efforts in global preparedness for disease outbreaks. While the overall genomic collection of the discovered viral pathogens continues to grow, simultaneously, the pathogens keep on evolving and there are both new emergence as well as reemergence of infectious diseases. In particular, the evolvability of RNA viruses, given their high mutation rates and large population sizes, allows such evolution to take place in actual timescales of human observation. The field of phylodynamics, therefore, seeks to combine data from both phylogenetics and epidemiological dynamics (Grenfell et al., 2004). This combined approach can allow more effective prediction of the epidemiological impact of newly emerged or evolved viruses. Further, Holmes and Grenfell (2009) suggest simultaneous collection of data of different types such as (a) spatiotemporal dynamics of the disease , (b) viral genome sequences, (c) contact networks of susceptible host individuals, and (d) the immune history of the individuals in contact networks, which, taken together, "are key for understanding both the dynamics of epidemic spread and the evolutionary pressures that shape virus diversity." An example of such integrative analysis is given in Cottam et al. (2008).

Clearly, such integrative analysis has the hallmark of Big Data Analytics where data have large size, are of diverse types, and generated dynamically. By analyzing possible interactions between the viral and immune dynamics within the infected hosts, one can determine the trajectory of pathogenic adaptations in real time. Algorithms to model normal immunologic profiles in the population can help in both outbreak prediction and detection (Pyne et al., 2009, 2014; Ray and Pyne, 2012). It can provide critical insights into the spatiotemporal spread of specific infections in given populations, help in estimation of R_0, and inferring viral phenotypes (e.g., virulence, transmissibility) (Volz et al., 2013). Under different phylodynamic processes, the viral phylogenetic tree shapes assume different spatiotemporal patterns depending on the strength of the immune-driven selection (Grenfell et al., 2004). Finally, phylodynamics also allows viral diversity to be studied statistically using the coalescent theory, which links phylogenetic structures with ecological processes, often to reconstruct demographic histories of infected individuals, or to estimate the parameters of infectious disease dynamics (Pybus and Rambaut, 2009; Suchard and Rambaut, 2009). For this purpose, parallel algorithms and high-performance computing have been used to support computationally intensive Bayesian frameworks to run on large data sets, e.g., Suchard and Rambaut (2009) and Ayres (2012). New platforms like OutbreakTools provide special formats to represent epidemiologic and sequence data together to allow statistical analysis, simulation, and visualization (Jombart et al., 2014). Owing to the HTS platforms, presently more than 16,000 human and avian isolates have been completely sequenced in the Influenza Genome Sequencing Project. In the near future, programs for discovery of pathogens and prediction of their transmission dynamics in host populations thus clearly seem to belong to the field of big data analytics.

6 HIGH-PERFORMANCE SYNTHETIC INFORMATION ENVIRONMENTS AND TOOLS

As discussed in Section 2, the key components of the the network-based SIR models of epidemic spread are the contact network and the disease spread model. The latter was discussed in Section 5. We now discuss how contact networks are modeled. These are very complex to analyze and simulate, and we discuss HPC tools for using such models.

6.1 Synthetic Networks for Epidemiology

Real data for large-scale social contact networks are not easily available, at least in the public domain, because of privacy and proprietary issues. As discussed earlier, a common approach to deal with lack of realistic data is to use simplified stochastic models, which match aggregate properties, e.g., degree distribution (Chung and Lu, 2002; Newman, 2003). However, as argued by Li et al. (2004), and many other researchers subsequently, there are significant limitations to such simplified models. We describe first-principles-based methods for generating synthetic regional and national scale social contact networks (Barrett et al., 2005; Eubank et al., 2004b). Unlike simple random graph techniques, these methods use real world data sources and combine them with behavioral and social theories to synthesize networks. This work builds on the TRANSIMS modeling environment (Barrett et al., 2000, 2001; Beckman et al., 1996) and has since been extended (Barrett et al., 2009). This approach integrates over a dozen public and commercial data sets and involves four broad steps discussed below; see Barrett et al. (2005, 2009); Eubank et al. (2004b) for more details, and a discussion on structural differences between synthetic networks and simple random networks.

The first step involves the creation of a synthetic urban population by integrating a variety of public and commercial data sets, while preserving their privacy and maintaining statistical indistinguishability. The synthetic population is a set of synthetic people and households, geographically located, each associated with demographic variables drawn from any of the demographics available in the input data set. Each synthetic individual is created in the context of a household with other synthetic individuals, and each household is located geographically. The second step involves the construction of a set of activity templates for households, based on several thousand responses to an activity or time-use survey. These activity templates include the sort of activities each household member performs and the time of day they are performed. Locations are selected for all activities, following capacity constraints, and models from transportation literature. Detailed route plans are assigned to individuals in the third step, based on the locations where activities are performed and the road network that connects the locations. Finally, detailed movement patterns are constructed using a cellular automata-based microsimulation for individuals over the transportation infrastructure.

The resulting synthetic population has a spatial resolution of few meters and a temporal resolution of a minute, across a large urban region. This information can be captured in the form of a *dynamic social contact network*, represented by a vertex and edge labeled bipartite graph G_{PL}, where P is the set of people and L is the set of locations. If a person $p \in P$ visits a location $\ell \in L$, there is an edge $(p, \ell, label) \in E(G_{PL})$ between them, where *label* is a record of the type of activity of the visit and its start and end times. Note that it is impossible to build such a network for any region large enough to be epidemiologically significant solely by collecting field data, although such data can be incorporated into the synthetic population creation process. The use of generative models to build such networks is a unique feature of this work.

Recently, researchers have included other forms of data and information to extend the basic methodology described above. Examples include (i) using information from airline data to construct network-based representation of cities across the globe—in this each node corresponds to a city and edge corresponds to number of travelers that go between the two cities as measured by air transport data (Colizza et al., 2006); (ii) representation of smaller subnetworks (aka micronetworks), using either survey data or data collected using sensors (Mossong et al., 2008; Salathé et al., 2010); and (iii) use of LandScan data in conjunction with census and other sources of population information to construct resolved networks that are not as accurate but can be constructed easily for several cities as well as countries (Ferguson et al., 2005; Longini et al., 2005).

6.2 Individual and Collective Behaviors

A primary goal of an epidemiologist is to control the spread of infectious disease through the application of interventions, guided by public policy. Policy-based interventions induce a behavioral change in individuals. At the same time, individuals self-impose behavioral changes in response to their perception of the current state of disease incidence. Both of these modifying factors imply that the underlying social network is constantly changing. In fact, individual behaviors, public policies, disease dynamics, and the social contact networks that they generate interact and coevolve as the outbreak progresses. The recent SARS epidemic served as an excellent example of how individual behavior as well as public policies played an important role in changing the social network.

Health scientists have developed verbal or conceptual behavioral models (Bandura, 1986; Becker, 1974) to understand the role of behaviors in public health. These models have played an important role in understanding behaviors and its relationship with diseases and maintaining a healthy life. These include the *Health Belief Model (HBM)* (Becker, 1974), models based on the Social Cognitive Theory (SCT) (Bandura, 1986), and the Social Ecological Model (SET). Verbal social and behavioral theories have proven useful in improving public health—but are often informal. These informal models need to be represented as formal computer procedures when using computer models to study epidemics. This turns out to be challenging. The role of information is critical in how such models are developed. We have investigated behavioral models based on (i) role of information (kind of information: space, time, and networked; type of information: local or global; and completeness of information), (ii) computational considerations (efficiency of encoding, computational resources required expressiveness of the language), and (iii) scale of behavior: individual, collective, and institutional, and it is quite demanding to identify the data necessary to instantiate *in silico* behavioral models. Recent advances in social media, computational turks, online games, online surveys, and digital traces are potentially exciting new ways to make progress in this direction. Of course availability of these data sets poses new kinds of questions, including (i) design to elicit truthful behavior, (ii) biases in data due to the demographics of participants, and (iii) translating behaviors in virtual world to the real world. See Funk et al. (2009), Salathé et al. (2012), and Barrett et al. (2010) for recent discussion on this emerging topic.

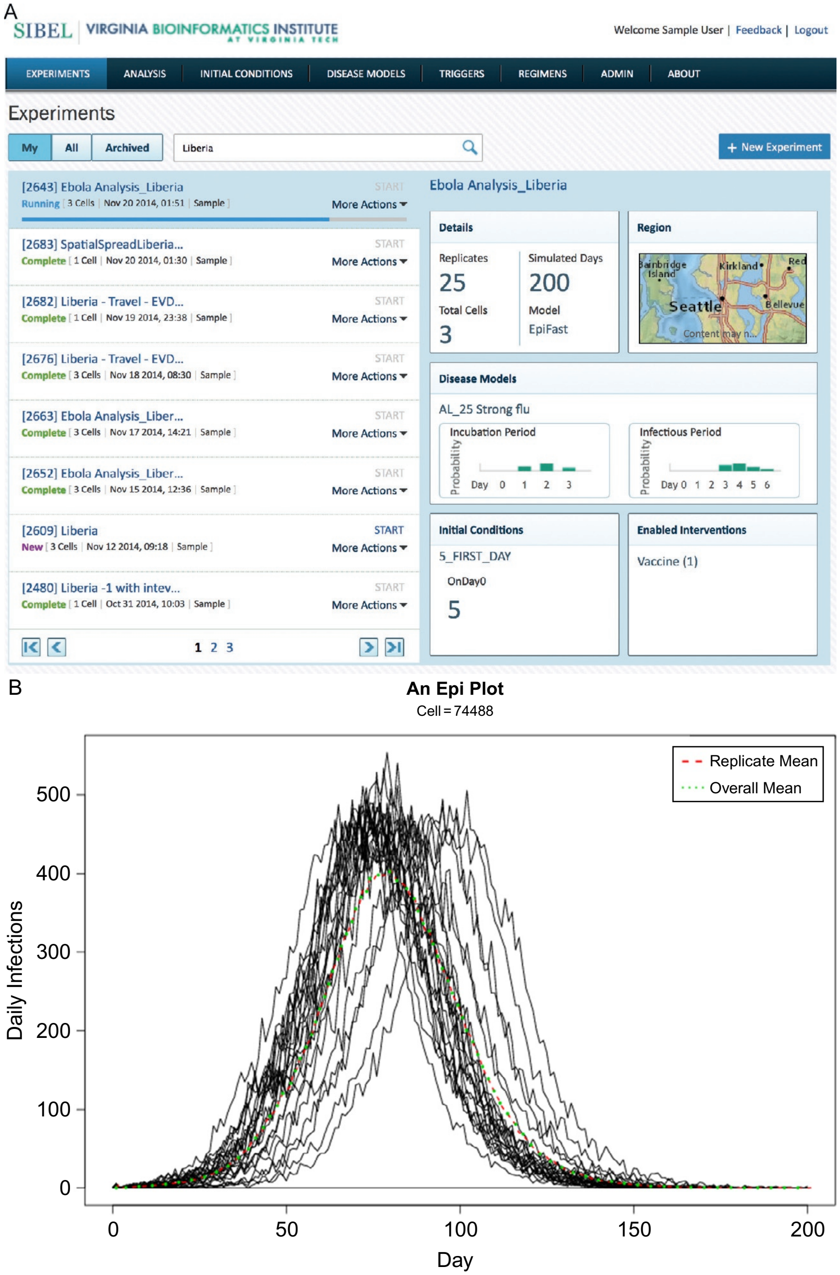
6.3 High-Performance Computing Tools

As discussed earlier in Section 3, phase-space properties of epidemic models have been analyzed mathematically for a small class of random and regular graph models. Computing disease dynamics over complex social contact networks is a computationally challenging problem, motivating the need for efficient simulations to calculate the spatiotemporal disease propagation. One of the first such simulations was EpiSims (Eubank, 2002; Eubank et al., 2004b; Mniszewski et al., 2008), which had a powerful modeling environment, but was not able to scale to large networks very easily. We have developed three different tools, Episimdemics (Barrett et al., 2008; Bisset et al., 2009b), EpiFast (Bisset et al., 2009a), and Indemics (Bisset et al., 2010a,b, 2011; Deodhar et al., 2012; Ma et al., 2011a,b)— these provide a trade-off between computation speed, model realism and sophistication, and ease of introducing new behavior and interventions. All three can be executed on traditional distributed memory clusters of varying sizes. EpiSimdemics is quite general and flexible, at the cost of decreased computational speed. It is an interaction-based, highly resolved modeling and simulation system for representing and reasoning about contagion diffusion across large (300 million nodes and 1.5 billion edges) networks. EpiFast (Bisset et al., 2009a) is a reaction-diffusion model that runs on an explicit social networks. It is based on a client–serve bulk synchronous computing paradigm and runs extremely fast—a single run on a network with 10 million nodes and 500 million edges take about 40 s on a 40-core machine. Indemics is an extension to EpiFast that integrates database computing to enable the most general set of interventions that are the easiest to create, but at the loss of computational speed (Bisset et al., 2010a,b, 2011; Deodhar et al., 2012; Ma et al., 2011a,b). This is an active area, and some of the recent work includes Perumalla and Seal (2011), Carley et al. (2006), Parker and Epstein (2012), Eubank (2002), Ferguson et al. (2006), Longini et al. (2005), Germann et al. (2006), and Chao et al. (2010).

6.4 Decision Support Environments

The epidemiological modeling tools described above are capable of providing very detailed information on spatiotemporal disease dynamics. The size and scale of the data and the expertise required to use the simulations demand a user friendly environment that provides an easy way to set up experiments and analyze the results. Recently, a number of visual analytics tools have been developed to support epidemiological research (see Livnat et al., 2010, 2012). We have built a tool called the SIBEL.[6] See http://isisdemo.vbi.vt.edu/didactic/isis.html for a demo version of SIBEL that allows a user to set up detailed factorial experiments (see Fig. 7). Using a simple interface to an underlying digital library, a user can choose from among many preconstructed instances: (i) a social contact network; (ii) a within-host disease progression model; and (iii) a set of interventions. Each intervention requires additional details such as compliance level, subpopulations to which the interventions are applied and intervention triggers. An experiment consists of sweeping one or more parameters across a user-specified range of values. After setting up the experiment, the user is provided access to the results of the simulations. A set of basic analyses are performed automatically and the results are displayed. The standard plots and epidemic curves provide very detailed information about the epidemic. Additional information such as the spatiotemporal dynamics and disease dendrogram (how the disease moved over the social network) is also available. A key aspect of this tool is its simplicity—we can train public health analysts to make effective use of the system in about 3 h. Several other groups are actively developing similar systems. They include: (i) The Biosurveillance Ecosystem (BSV) being developed by DTRA; (ii) The BARD model repository at the Los Alamos National Laboratory; (iii) The Texas Pandemic toolkit being developed at the University of Texas, Austin, http://flu.tacc.utexas.edu/; (iv) The MIDAS funded Apollo project at the University of Pittsburgh and the framework at RTI; (v) The FRED modeling framework at the University of Pittsburgh; and (vi) The EpiC framework being developed by MoBs laboratory at Northeastern University.

[6] Earlier versions of SIBEL were called DIDACTIC and ISIS. The change in the name reflects new functionality.



Liberia

[2643] Ebola Analysis_Liberia
Running [3 Cells | Nov 20 2014, 01:51 | Sample] START More Actions

[2683] SpatialSpreadLiberia...
Complete [1 Cell | Nov 20 2014, 01:30 | Sample] START More Actions

[2682] Liberia - Travel - EVD...
Complete [1 Cell | Nov 19 2014, 23:38 | Sample] START More Actions

[2676] Liberia - Travel - EVD...
Complete [3 Cells | Nov 18 2014, 08:30 | Sample] START More Actions

[2663] Ebola Analysis_Liber...
Complete [3 Cells | Nov 17 2014, 14:21 | Sample] START More Actions

[2652] Ebola Analysis_Liber...
Complete [3 Cells | Nov 15 2014, 12:36 | Sample] START More Actions

[2609] Liberia
New [3 Cells | Nov 12 2014, 09:18 | Sample] START More Actions

[2480] Liberia -1 with intev...
Complete [1 Cell | Oct 31 2014, 10:03 | Sample] START More Actions

1 2 3

Ebola Analysis_Liberia

Details

Replicates 25

Simulated Days 200

Total Cells 3

Model EpiFast

Region

Disease Models

AL_25 Strong flu

Initial Conditions

5_FIRST_DAY

OnDay0

5

Enabled Interventions

Vaccine (1)

FIGURE 7 SIBEL interface (a). Example simulated epidemic curves (b).

7 SUMMARY

In summary, computational epidemiology is a new and exciting multidisciplinary area with significant challenges of large data and high-performance computing. Recent developments in high-performance pervasive computing have created new opportunities for collecting, integrating, analyzing, and accessing information related to large sociotechnical networks. The advances in network and information science that build on this new capability provide entirely new ways for reasoning and controlling epidemics. Together, they enhance our ability to formulate, analyze, and realize novel public policies pertaining to these complex systems.

There are many other important issues in epidemiology that we have not addressed here. An important area is that of vector borne diseases and zoonotic diseases. Developing computational methods to understand and control these two classes present unique new challenges that one does not have to encounter while studying human–human transmission. Another area where similar ideas have been applied is the epidemiology of noncommunicable diseases, such as obesity and diabetes, and addictions such as smoking. Further, there is a broader class of reaction–diffusion models generalizing the SIR/SIS models, which has been applied to diverse phenomena in economics, sociology, viral marketing, and political science; we refer the readers to Marathe and Vullikanti (2013) for more pointers to these topics.

ACKNOWLEDGMENTS

We thank our external collaborators and members of the Network Dynamics and Simulation Science Laboratory (NDSSL) for their suggestions and comments. S.P. was supported by Ramalingaswami-DBT Fellowship, DST-CMS, and MoS&PI India (Project No. SR/SA/MS:516/07 dated 21/04/2008). The work of A.K.S.V. and M.V.M. has been partially supported by National Science Foundation: NSF NetSE Grant CNS-1011769, NSF SDCI Grant OCI-1032677, Defense Threat Reduction Agency Grant HDTRA1-11-1-0016, DTRA CNIMS Contract HDTRA1-11-D-0016-0001, National Institute of Health Midas Grant 2U01GM070694-09, and by the Intelligence Advanced Research Projects Activity (IARPA) via the US Department of Interior (DoI) National Business Center (NBC) Contract No. D12PC000337. The US government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. 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, or the US government.

REFERENCES

Allen, L.Z., Ishoey, T., Novotny, M.A., McLean, J.S., Lasken, R.S., Williamson, S.J., 2011. Single virus genomics: a new tool for virus discovery. PLoS One 6 (3), e17722. http://dx.doi.org/10.1371/journal.pone.0017722.

Alon, N., Benjamini, I., Stacey, A., 2004. Percolation on finite graphs and isoperimetric inequalities. Ann. Probab. 32 (3), 1727–1745.

Aspnes, J., Chang, K.L., Yampolskiy, A., 2006. Inoculation strategies for victims of viruses and the sum-of-squares partition problem. J. Comput. Syst. Sci. 72 (6), 1077–1093.

Ayres, D.L., 2012. Beagle: an application programming interface and high-performance computing library for statistical phylogenetics. Syst. Biol. 61 (1), 170–173.

Bandura, A., 1986. Social Foundations of Thought and Action: A Social Cognitive Theory. Englewood Cliffs, NJ, US: Prentice-Hall, Inc.

Barrett, C.L., Beckman, R., Berkbigler, K., Bisset, K., Bush, B., Campbell, K., Eubank, S., Henson, K., Hurford, J., Kubicek, D., Marathe, M., Romero, P., Smith, J., Smith, L., Speckman, P., Stretz, P., Thayer, G., Eeckhout, E., Williams, M., 2000. TRANSIMS (TRansportation ANalysis SIMulation System). Los Alamos National Laboratory. Technical report no. LA-UR-00-1725.

Barrett, C.L., Jacob, R., Marathe, M., 2001. Formal language constrained path problems. SIAM J. Comput. 30 (3), 809–837.

Barrett, C., Smith, J.P., Eubank, S., 2005. Modern Epidemiology Modeling. Scientific American.

Barrett, C., et al., 2008. Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In: Proceedings of the ACM/IEEE Conference on High Performance Computing (SC).

Barrett, C., et al., 2009. Generation and analysis of large synthetic social contact networks. In: Proceedings of the Winter Simulation Conference, pp. 1003–1014.

Barrett, C., Bisset, K., Leidig, J., Marathe, A., Marathe, M., 2010. An integrated modeling environment to study the co-evolution of networks, individual behavior and epidemics. AI Mag. 31 (1), 75–87.

Barzon, L., et al., 2011. Applications of next generation sequencing technologies to diagnostic virology. Int. J. Mol. Sci. 12(11). 7861–7884.

Becker, M., (Ed.), 1974. The health belief model and personal health behavior. Health Education Monogr. 2 (no. 2), 324–508.

Beckman, R., Baggerly, K., McKay., M., 1996. Creating base-line populations. Transport. Res. A Policy Pract. 30, 415–429.

Bisset, K., Chen, J., Feng, X., Vullikanti, A. and Marathe, M. 2009a. EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In: Proceedings of 23rd ACM International Conference on Supercomputing (ICS′ 09). ACM Press, New York.

Bisset, K., Feng, X., Marathe, M., Yardi, S., 2009b. Modeling interaction between individuals, social networks and public policy to support public health epidemiology. In: Proceedings of the 2009 Winter Simulation Conference, pp. 2020–2031.

Bisset, K., Chen, J., Feng, X., Ma, Y., Marathe, M., 2010a. Indemics: an interactive data intensive framework for high performance epidemic simulation. In: Proceedings of the 24th ACM International Conference on Supercomputing, pp. 233–242.

Bisset, K., et al., 2010b, Indemics: an interactive data intensive framework for high performance epidemic simulation. In: Proceedings of the International Conference on Supercomputing.

Bisset, K., et al., 2011. Interaction-based hpc modeling of social, biological, and economic contagions over large networks. In: Proceedings of Winter Simulation Conference (WSC), pp. 2933–2947.

Blainey, P.C., 2013. The future is now: single-cell genomics of bacteria and archea. FEMS Microbiol. Rev. 37 (3), 407–427.

Borgs, C., Chayes, J.T., Ganesh, A., Saberi, A., 2010. How to distribute antidotes to control epidemics. Random Struct. Algorithms 37, 204–222.

Borozan, I., et al., 2012. CaPSID: a bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. BMC Bioinformatics 13 (206), 1–11.

Brauer, F., van den Driessche, P., Wu, J. (Eds.), 2008. Mathematical Epidemiology. Lecture Notes in Mathematics 1945. Springer Verlag, Berlin, Heidelberg, New York.

Buczak, A.L., Babin, S.M., Feighner, B.H., Koshute, P.T., Lewis, S.H., 2013. Predictive modeling of emerging infections. In: Viral Infections and Global Change, Singh, S.K. (Ed.), John Wiley & Sons, Inc., Hoboken, NJ.

Carley, K.M., Fridsma, D.B., Casman, E., Yahja, A., Altman, N., Chen, L.C., Kaminsky, B., Nave, D., 2006. Biowar: scalable agent-based model of bioattacks. IEEE Trans. Syst. Man Cybernet. A 36 (2), 252–265.

Chakraborty, P., Khadivi, P., Lewis, B., Mahendiran, A., Chen, J., Butler, P., Nsoesie, E., Mekaru, S., Brownstein, J., Marathe, M., Ramakrishnan, N., 2014a, Forecasting a moving target: ensemble models for ILI case count predictions. In: Proceedings of the SIAM International Conference on Data Mining (SDM).

Chakraborty, P., Khadivi, P., Lewis, B., Mahendiran, A., Chen, J., Butler, P., Nsoesie, E.O., Mekaru, S.R., Brownstein, J.S., Marathe, M.V., Ramakrishnan, N., 2014b. Forecasting a moving target: ensemble models for ILI case count predictions. In: Proceedings of the 2014 SIAM International Conference on Data Mining, 28 April 2014, pp. 262–270. http://dx.doi.org/10.1137/1.9781611973440.30.

Chao, D.L., Halloran, M.E., Obenchain, V., Longini Jr., I.M., 2010. FluTE, a publicly available stochastic influenza epidemic simulation model. PLoS Comput. Biol. 6 (1), e1000656.

Chen, E.C., Miller, S.A., DeRisi, J.L., Chiu, C.Y., 2011. Using a pan-viral microarray assay (virochip) to screen clinical samples for viral pathogens. J. Vis. Exp. 50. pii: 2536. http://dx.doi.org/10.3791/2536.

Chiu, C.Y., 2013. Viral pathogen discovery. Curr. Opin. Microbiol. 16, 468–478.

Chung, F., Lu, L., 2002. Connected components in random graphs with given degree sequences. Ann. Combinatorics 6, 125–145.

Coker, R.J., et al., 2011. Towards a conceptual framework to support one health research for policy on emerging zoonoses. Lancet Infect. Dis. 11, 326–331.

Colizza, V., Barrat, A., Barthelemy, M., Vespignani, A., 2006. The role of the airline transportation network in the prediction and predictability of global epidemics. PNAS 103 (7), 2015–2020. doi:10.1073/pnas.0510525103. http://www.pnas.org/content/103/7/2015.full.pdf+html, http://www.pnas.org/content/103/7/2015.abstract.

Cooper, G.E., et al., 2006. Bayesian methods for diagnosing outbreaks. In: Wagner, M.M., Moore, A.W., Aryel, R.M. (Eds.), Handbook of Biosurveillance. Elsevier, New York City, NY.

Cottam, E.M., et al., 2008. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proc. R. Soc. Lond. B. 275 (1637), 887–895.

Deodhar, S., Bisset, K., Chen, J., Ma, Y., Marathe, M.V., 2012. Enhancing software capability through integration of distinct software in epidemiological systems. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pp. 171–180.

Dimitrov, N.B., Meyers, L.A., 2010. Mathematical approaches to infectious disease prediction and control. In: Hasenbein, J.J. (Ed.), INFORMS TutORials in Operations Research 7, pp. 1–25.

Easley, D., Kleinberg, J., 2010. Networks, Crowds and Markets: Reasoning About A Highly Connected World. Cambridge University Press, New York, NY.

Eubank, S., 2002. Scalable, efficient epidemiological simulation. In: SAC '02: Proceedings of the 2002 ACM Symposium on Applied Computing. ACM, New York, NY, pp. 139–145.

Eubank, S., Kumar, V.A., Marathe, M., Srinivasan, A., Wang, N., 2004a. Structural and algorithmic aspects of massive social networks. In: ACM Symposium on Discrete Algorithms (SODA).

Eubank, S., Guclu, H., Anil Kumar, V.S., Marathe, M., Srinivasan, A., Toroczkai, Z. Wang, N., 2004b. Modelling disease outbreaks in realistic urban social networks. Nature 429, 180–184.

Ferguson, N.M., et al., 2005. Strategies for containing an emerging influenza pandemic in southeast Asia. Nature 437, 209–214.

Ferguson, N.M., et al., 2006. Strategies for mitigating an influenza pandemic. Nature 442, 448–452.

Finn, R., Clements, J., Eddy, S., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37.

Firth, C., Lipkin, W.I., 2013. The genomics of emerging pathogens. Annu. Rev. Genomics Hum. Genet. 14, 281–300.

Fox, G., Mani, D.R., Pyne, S., 2013. Parallel deterministic annealing clustering and its application to LC-MS data analysis. In: Proceedings of the IEEE International Conference on Big Data.

Funk, S., Gilad, E., Watkins, C., Jansen, V.A.A., 2009. The spread of awareness and its impact on epidemic outbreaks. PNAS 106 (16), 6872–6877.

Funk, S., Salathé, M., Jansen, V., 2010. Modelling the influence of human behaviour on the spread of infectious diseases: a review. J. R. Soc. Interface 7, 1247–1256. doi:10.1098/rsif.2010.0142.

Ganesh, A., Massoulie, L., Towsley, D., 2005. The effect of network topology on the spread of epidemics. Proc. INFOCOM. 2, 1455–1466.

Gardner, S.N., et al., 2010. A microbial detection array (MDA) for viral and bacterial detection. BMC Genomics 11, 668.

Germann, T.C., Kadau, K., Longini, I.M., Macken, C.A., 2006. Mitigation strategies for pandemic influenza in the united states. Proc. Natl. Acad. Sci. U.S.A. 103 (15), 5935–5940. doi:10.1073/pnas.0601266103.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014.

Gomez-Rodriguez, M., Leskovec, J., Krause, A., 2010. Inferring networks of diffusion and influence. In: Proceedings of the 16th ACM KDD.

Grenfell, B.T., et al., 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303, 327–332.

Hall, I.M., Gani, R., Hughes, H.E., Leach, S., 2007. Real-time epidemic forecasting for pandemic influenza. Epidemiol. Infect. 135 (3), 372–385.

Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P., Beutels, P., 2012. Modeling infectious disease parameters based on serological and social contact data. A modern statistical perspective. Statistics for Biology and Health. Springer, New York, NY. doi:10.1007/978-1-4614-4072-7.

Holmes, E.C., 2009. The Evolution and Emergence of RNA Viruses. Oxford University Press, New York, NY.

Holmes, E.C., Grenfell, B.T., 2009. Discovering the phylodynamics of RNA viruses. PLoS Comput. Biol. 5 (10), e1000505.

Jombart, T., et al., 2014. OutbreakTools: new platform for disease outbreak analysis using the R software. Epidemics, 7, 28–34.

Jones, K.E., et al., 2008. Global trends in emerging infectious diseases. Nature 451, 990–993.

King, D.A., et al., 2006. Infectious diseases: preparing for the future. Science 313, 1392–1393.

Kostic, A., et al., 2011. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat. Biotechnol. 29, 393–396.

Last, J., 2001. A Dictionary of Epidemiology, fourth ed. Oxford University Press, New York.

Li, L., Alderson, D., Willinger, W., Doyle, J., 2004. A first principles approach to understanding the internet's router level topology. In: Proceedings of the ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM), pp. 3–14.

Li, W.D., et al., 2005. Bats are natural reservoirs of SARS-like coronaviruses. Science 310 (5748), 676–679.

Lipkin, W.I., 2009. Microbe hunting in the 21st century. Proc. Natl. Acad. Sci. U.S.A. 106 (1) 6–7.

Lipkin, W.I., 2013. The changing face of pathogen discovery and surveillance. Nat. Rev. Microbiol. 11, 133–141.

Lipsitch, M., Cohen, T., Cooper, B., Robins, J.M., Ma, S., James, L., Gopalakrishna, G., Chew, S., Tan, C.C., Samore, M.H., Fisman, D., Murray, M., 2003. Transmission dynamics and control of severe acute respiratory syndrome. Science 300, 1966–1970.

Lipsitch, M., et al., 2009. Managing and reducing uncertainty in an emerging influenza pandemic. N. Engl. J. Med. 361 (2), 112–115. http://www.nejm.org/doi/pdf/10.1056/NEJMp0904380, http://www.nejm.org/doi/full/10.1056/NEJMp0904380.

Livnat, Y., Gesteland, P., Benuzillo, J., Pettey, W., Bolton, D., Drews, F., Kramer, H., Samore, M., 2010. Epinome—a novel workbench for epidemic investigation and analysis of search strategies in public health practice. Proc. Annu. Am. Med. Inform. Assoc. Sympos. 647–651.

Livnat, Y., Rhyne, T., Samore, M., 2012. Epinome: A visual-analytics workbench for epidemiology data. IEEE Comput. Graph. Appl. 32 (2), 89–95.

Loman, N.J., et al., 2012. Performance comparison of bench-top high-throughput sequencing platforms. Nat. Biotechnol. 30, 434–439.

Longini, I.M., et al., 2005. Containing pandemic influenza at the source. Science 309, 1083–1087.

Ma, Y., Bisset, K.R., Chen, J., Deodhar, S., Marathe, M.V., 2011a, Formal specification and experimental analysis of an interactive epidemic simulation framework. In: Proceedings of the 2011 International Workshop on Extreme Scale Computing Application Enablement—Modeling and Tools (ESCAPE).

Ma, Y., Bisset, K.R., Chen, J., Deodhar, S., Marathe, M.V., 2011b, Efficient implementation of complex interventions in large scale epidemic simulations. In: Proceedings of the Winter Simulation Conference.

Marathe, M., Vullikanti, A., 2013. Computational epidemiology. Commun. ACM. 56 (7), 88–96.

Metzker, M.L., 2009. Sequencing technologies—the next generation. Nat. Rev. Genet. 11 (11), 31–46.

Mniszewski, S.M., Del Valle, S.Y., Stroud, P.D., Riese, J.M., Sydoriak, S.J., 2008. EpiSimS simulation of a multi-component strategy for pandemic influenza. In: Proceedings of the 2008 Spring Simulation Multiconference, SpringSim '08. Society for Computer Simulation International, San Diego, CA, pp. 556–563. http://portal.acm.org/citation.cfm?id=1400549.1400636.

Mokili, J.L., et al., 2012. Metagenomics and future perspectives in virus discovery. Curr. Opin. Virol. 2, 63–77.

Morse, S.S., et al., 2012. Prediction and prevention of the next pandemic zoonosis. Lancet 380, 1956–1965.

Mortveit, H.S., Reidys, C.M., 2007. An Introduction to Sequential Dynamical Systems, Universitext. Springer Verlag, New York.

Mossong, J., et al., 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 5 (3), e74. http://dx.doi.org/10.1371/journal.pmed.0050074.

Naccache, S.N., et al., 2014. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res. 24, 1180–1192.

Naeem, R., Rashid, M., Pain, A., 2013. READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. Bioinformatics 29 (3), 391–392.

Neil, D., Moore, A., Cooper, G., 2005. A bayesian scan statistic for spatial cluster detection. In: Proceedings of the National Syndromic Surveillance Conference.

Netrapalli, P., Sanghavi, S., 2012. Learning the graph of epidemic cascades. In: ACM SIGMETRICS.

Newman, M.E.J., 2003. The structure and function of complex networks. SIAM Rev. 45, 167–256.

Nishiura, H., 2011. Real-time forecasting of an epidemic using a discrete time stochastic model: a case study of pandemic influenza (H1N1-2009). BioMed. Eng. Online 10 (1), 15.

Nsoesie, E.O., Beckman, R., Marathe, M., Lewis, B., 2011. Prediction of an epidemic curve: a supervised classification approach. Stat. Commun. Infect. Dis. 3 (1), 5.

Nsoesie, E.O., Brownstein, J.S., Ramakrishnan, N., Marathe, M., 2013. A systematic review of studies on forecasting the dynamics of influenza outbreaks. Influenza Other Respir. Viruses 8 (3), 309–316.

Ohkusa, Y., Sugawara, T., Taniguchi, K., Okabe, N., 2011. Real-time estimation and prediction for pandemic A/H1N1(2009) in Japan. J. Infect. Chemother. 17 (4), 468–472.

Palacios, G., et al., 2007. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. Emerg. Infect. Dis. 13, 73–81.

Parker, J., Epstein, J.M., 2012. A distributed platform for global-scale agent-based models of disease transmission. ACM Trans. Model. Comput. Simulation 22 (1), Article No. 2.

Pastor-Satorras, R., Vespignani, A., 2001, Apr. Epidemic spreading in scale-free networks. Phys. Rev. Lett. 86, 3200–3203. http://link.aps.org/doi/10.1103/PhysRevLett.86.3200.

Patz, J.A., et al., 2004. Unhealthy landscapes: policy recommendation on land use change and infectious disease emergence. Environ. Health Perspect. 112 (10), 1092–1098.

Perumalla, K., Seal, S., 2011. Discrete event modeling and massively parallel execution of epidemic outbreak phenomena. Simulation 88 (7), 768–783.

Pybus, O.G., Rambaut, A., 2009. Evolutionary analysis of the dynamics of viral infectious disease. Nat. Rev. Genet. 10, 540–550.

Pyne, S., et al., 2009. Automated high-dimensional flow cytometric data analysis. Proc. Natl. Acad. Sci. U.S.A. 106.

Pyne, S., et al., 2014. Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. PLoS One. 9, e100334.

Ray, S., Pyne, S., 2012. A computational framework to emulate the human perspective in flow cytometric data analysis. PLoS One. 7, e35693.

Riesenfeld, C.S., 2004. Metagenomics: genomic analysis of microbial communities. Annu. Rev. Genet. 38, 525–552.

Salathé, M., et al., 2010. A high-resolution human contact network for infectious disease transmission. PNAS 107 (51), 22020–22025. ISSN 1091-6490. http://dx.doi.org/10.1073/pnas.1009094108.

Salathé, M., et al., 2012. Digital epidemiology. PLoS Comput. Biol. 8 (7), e1002616.

Schwind, J.S., et al., 2014. Capacity building efforts and perceptions for wildlife surveillance to detect zoonotic pathogens: comparing stakeholder perspectives. BMC Public Health 14, 684.

Shah, D., Zaman, T., 2010. Rumors in a network: who is the culprit? ACM SIGMETRICS.

Shaman, J., Karspeck, A., 2012. Forecasting seasonal outbreaks of influenza. Proc. Natl. Acad. Sci. U.S.A. 109 (50), 20425–20430. http://www.pnas.org/content/109/50/20425.full.pdf+html, http://www.pnas.org/content/109/50/20425.abstract.

Shaman, J., Goldstein, E., Lipsitch, M., 2010a. Absolute humidity and pandemic versus epidemic influenza. Am. J. Epidemiol. 173 (2), 127–135.

Shaman, J., Pitzer, V.E., Viboud, C., Grenfell, B.T., Lipsitch, M., 2010b. Absolute humidity and the seasonal onset of influenza in the continental United States.. PLoS Biol. 8 (2), e1000316.

Suchard, M.A., Rambaut, A., 2009. Many-core algorithms for statistical phylogenetics. Bioinformatics 25, 1370–1376.

Taylor, L.H., et al., 2001. Risk factors for human disease emergence. Philos. Trans. R. Soc. Lond. B Biol. Sci. 356, 983–989.

Tizzoni, M., Bajardi, P., Poletto, C., Ramasco, J., Balcan, D., Goncalves, B., Perra, N., Colizza, V., Vespignani, A., 2012. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. BMC Med. 10 (1), 165. ISSN 1741-7015. http://www.biomedcentral.com/1741-7015/10/165.

Vaidyanathan, G., 2011. Virus hunters: catching bugs in the field. Cell 147, 1209–1211.

Volz, E.M., et al., 2013. Viral phylodynamics. PLoS Comput. Biol. 9, e1002947.

Wang, Y., Chakrabarti, D., Wang, C., Faloutsos, C., 2003. Epidemic spreading in real networks: an eigenvalue viewpoint. In: Proceedings of SRDS.