DOI:10.1145/2483852.2483871

The challenge of developing and using computer models to understand and control the diffusion of disease through populations.

BY MADHAV MARATHE AND ANIL KUMAR S. VULLIKANTI

Computational Epidemiology

AN EPIDEMIC IS said to arise in a community or region when cases of an illness or other health-related events occur in excess of normal expectancy. Epidemics are considered to have influenced significant historical events, including the plagues in Roman times and Middle Ages, the fall of the Han empire in the 3rd century in China, and the defeat of the Aztecs in the 1500s, due to a smallpox outbreak.[9] The 1918 flu pandemic in the U.S. was responsible for more deaths than those due to World War I. The last 50 years have seen epidemics caused by HIV/AIDS, SARS, and influenza-like illnesses. Despite significant medical advances, according to the World Health Organization (WHO), infectious diseases account for more than 13 million deaths a year.[44]

Societal interest in controlling outbreaks is probably just as old as the diseases themselves. Interestingly, it appears the Indians and Chinese knew the idea of variolation to control smallpox as early as the 8th century A.D. Epidemiology is a formal branch of science focusing on the study of space-time patterns of illness in a population and the factors that contribute to these patterns. It plays an essential role in public health by

key insights

- Controlling and responding to future pandemics will be challenging due to a number of emerging global trends including increased and denser urbanization, increased local as well as global travel, and a generally older and immuno-compromised population.
- Public health epidemiology is a complex system problem. Epidemics, social-contact networks, individual and collective behavior, and public policies coevolve during a pandemic—a system-level understanding must represent these components and their coevolution.
- Mathematical and computational models of social networks and epidemic spread and methods to analyze them are critical in public health epidemiology.
- Advances in computing, big data, and computational thinking have created entirely new opportunities to support real-time epidemiology.

ILLUSTRATION BY AS&K VISUAL SCIENCE, WWW.ASK-VISUAL.COM

aiming to understand the processes that lead to ill health as well as the evaluation of strategies designed to prevent illness and promote good health.

Computational epidemiology is an interdisciplinary area setting its sights on developing and using computer models to understand and control the spatiotemporal diffusion of disease through populations. The models may range from descriptive, for example, static estimates of correlations within large databases, to generative, for example, computing the spread of disease via person-to-person interactions through a large population. The disease may represent an actual infectious disease, or following the relatively recent use of this term, it may represent a more general reaction-diffusion process, such as the diffusion of innovation. The populations of interest depend on the disease, including humans, animals, plants, and computers. Similarly, the interactions that must be represented depend on the disease and the populations, including physical proximity for aerosol-borne disease, sexual contact for sexually transmitted diseases, and insect feeding patterns for mosquito-borne diseases.

In an editorial in *Science* (May 2009), Fineberg and Wilson persuasively argue for the role of science in policy as it pertains to real-time epidemiology.[26–28] They go on to identify five areas where science can support real-time epidemiology: pandemic risk; vulnerable populations; available interventions; implementation possibilities; and pitfalls and public understanding.

Computation can play a multifaceted role to support real-time epidemiology. The role of computation becomes all the more important in light of the fact that controlled experiments used to understand scientific phenomena are much harder and often impossible in epidemiology due to ethical and practical reasons. Figure 1 illustrates an end-to-end computationally oriented view to support real-time epidemiology including the five areas identified by Fineberg and Wilson. Computational models help in understanding the space-time dynamics of epidemics. Models and algorithms can be used to evaluate various intervention strategies, including pharmaceutical interventions such as vaccinations and anti-virals, as well as non-pharmaceutical interventions such as social distancing and school closures.

The role of individual behavior and public policies is critical in understanding and controlling epidemics—computational techniques provide a potentially powerful study tool.[24,27] Pervasive computational environments can provide real-time access to models, data, and expert opinion to analysts as an epidemic unfolds. It is important for computational methods and thinking to be developed within an inherently multidisciplinary setting. For example, inference methods will need to be developed for specific problems in conjunction with statisticians and biologists. The problem of vaccine production, allocation, and usage must be solved in conjunction with social, economic, and behavioral scientists to account for efficient use, keeping in mind social and economic factors. User interfaces and decision support environments would have to be developed in conjunction with psychologists and statisticians to avoid pitfalls in human judgment. Details omitted here because of space limitations can be found in supplemental material.[44]

Mathematical Modeling

The first mathematical model in epidemiology is credited to Bernoulli in 1760.[9] Using mathematical techniques, Bernoulli established that variolation 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.[9] The early 1900s saw seminal advances in mathematical epidemiology. Ross in 1911 found malaria spread due to mosquitoes, and developed a spatial model for the spread of malaria. 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 involving ordinary differential equations based on a mass-action model. Their work laid the modern foundations for mathematical epidemiology. (For details, see the sidebar "Modeling Epidemic Processes," supplementary information,[44] Brauer,[9] and Newman.[37])

Modeling Epidemic Processes

The simplest aggregate model is popularly known as the SIR model. A population of size N is divided into three states: susceptible (S), infective (I), and removed or recovered (R).

The following discrete time 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, $s(t) + i(t) + r(t) = 1$.

By the "complete mixing" assumption that each individual is in contact with everyone in the population, it can be shown that the following system of differential equations (known as the SIR model) describes the dynamics:

$$\frac{ds(t)}{dt} = -\beta s(t)\, i(t), \quad \frac{di(t)}{dt} = \beta s(t)\, i(t) - \gamma i(t), \quad \frac{dr(t)}{dt} = \gamma i(t).$$

One of the classic results in the SIR model is there is an epidemic that infects a large fraction of the population, if and only if $R_0 = \beta/\gamma > 1$; R_0 is known as the "reproductive number," and thus much of public health decision making is centered on controlling R_0.

Epidemics on Networks

Let $G(V, E)$ denote a contact graph on a population V—each edge $e = (u, v) \in E$ denotes that the individuals (also referred to as nodes) $u, v \in V$ come into contact. Let $N(v)$ denote the set of neighbors of v.

For the SIR model on the graph G, we have a dynamical process with each node being in 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 discrete version of the rate of infection in the sidebar "Modeling Epidemic Processes."

Let $\tau(u)$ denote the time that node u would remain in the infected state, and let $\tau = \max\{\tau(u) : u \in V\}$. If a node $v \in V$ gets infected at time t_v, it attempts to infect each susceptible neighbor u with probability $\beta((v, u), t - t_v)$ for $t = t_v + 1, \ldots, t_v + \tau(v)$. After $\tau(v)$ steps, node v switches to state R.

We let $I(t)$ denote the set of nodes that become infected at time t. The sequence $I(t)$, $t = 0, 1, \ldots$ along with the (random) subset of edges on which the infections spread, represent a disease outcome, also referred to as a *dendogram*.

The time series ($|I(t)|$, $t = 0, 1, \ldots$) is referred to as an epidemic curve corresponding to a stochastic outcome. The total number of infections for an outcome is given by $\Sigma_t\, |I(t)|$. The epidemic peak is the largest time t that maximizes $|I(t)|$. Thus, 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. A popular variant of the SIR model is the SIS model, in which an infected node switches to state S after the infectious duration.

Networked Epidemiology

The analysis in the sidebar "Modeling Epidemic Processes" involves a number of assumptions, chief among them being the complete mixing requirement, which is often unrealistic. Many researchers have extended this in numerous ways, including stochastic models, multiple compartments to represent various subpopulations, branching processes, and chain-binomial models, among others.[9,37] Here we will focus on a recent approach that is inherently computational in nature—*networked epidemiology* (see Figure 2). It seeks to understand the complicated interplay between the three components of computational epidemiology: individual behaviors of agents; unstructured; heterogeneous multiscale networks; and 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-

Figure 1. Elements of computational real-time epidemiology.

The process starts by developing models of synthetic social contact networks and within host disease progression using diverse datasets that include: surveys, census, social media, serological investigations, and disease surveillance. High-performance computer simulations are then used to study the dynamics aof disease propagation and the effects of various intervention strategies. The results are used by policymakers and analysts to formulate and evaluate various public policies as well as putative societal responses. All these models are refined, based on the simulation results and the policies being studied.

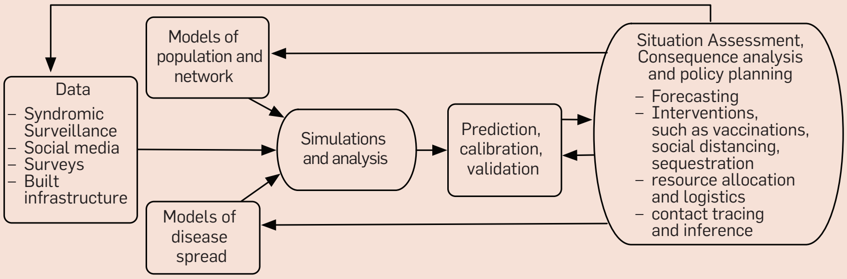
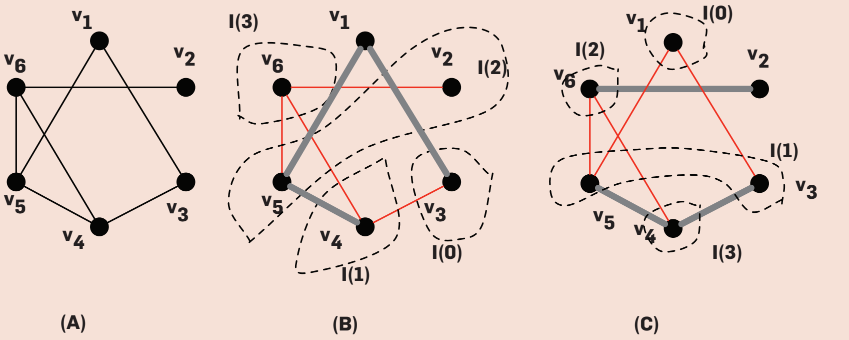

Models of population and network
Data
– Syndromic Surveillance
– Social media
– Surveys
– Built infrastructure
Simulations and analysis
Prediction, calibration, validation
Situation Assessment, Consequence analysis and policy planning
– Forecasting
– Interventions, such as vaccinations, social distancing, sequestration
– resource allocation and logistics
– contact tracing and inference
Models of disease spread

Figure 2. (a) Example showing a contact network on a population of size 6, represented by the set of nodes $\{v_1, v_2, v_3, v_4, v_5, v_6\}$. (b) An example of a dendogram on this contact graph, with the infected sets $I(t)$, $t = 0, 1, 2, 3$ as shown. The red edges represent the edges on which the infections spread. The infection starts at node v_3, and eventually all nodes, except v_1 get infected; the epicurve corresponding to this example is (1, 1, 2, 1), with the peak being at time 2. (c) Another possible dendogram on the same network, with the infection starting at v_1, where all nodes except v_2 get infected.



based epidemiology alters the types of questions posed, providing qualitatively different insights into disease dynamics and public health policies. It also allows policymakers to formulate and investigate potentially novel and context-specific interventions.

The sidebar "Epidemics on Networks" summarizes the underlying mathematical framework to represent epidemics over a network. A general framework to succinctly specify disease-local dynamical processes on networks is called the graphical discrete dynamical system (GDDS).[7] A GDDS $\mathcal{G}$ is defined as a tuple (G, F, π), where: (a) $G = (V, E)$ is the underlying contact network on a set V of nodes; (b) $F = \{f_v | v \in V\}$ is a set of local functions, (for example, such as the one capturing the localized SIR process)—one function for each node $v \in V$ on some fixed domain, where v computes its state by applying f_v on the states of its neighbors; and (c) a schedule $\mathcal{S}$ over V^* that specifies the order in which the states of the nodes are updated by applying the functions in F. One update of GDDS involves applying the local functions in the order specified by $\mathcal{S}$. Extensions of GDDS can be used to describe multitheory, multinetwork coevolving systems (referred to as CGDDS), wherein multiple networks and the local functions interact and coevolve in time endogenously.

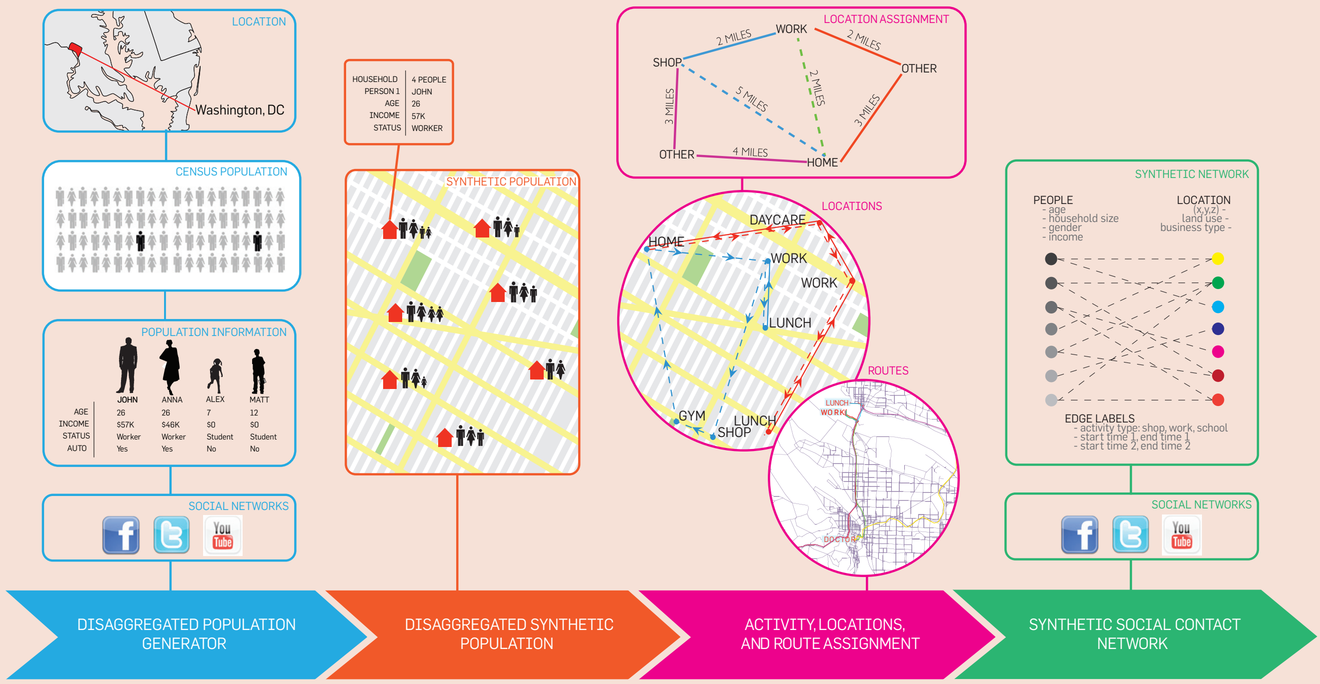
The configuration space of GDDS $\mathcal{G}$ can be conveniently viewed as a Markov chain $\mathcal{M}$. Each node in $\mathcal{M}$ is the state vector of the node states in G. A directed edge from node A and B in $\mathcal{M}$ with label p denotes the probability of transition from A to B. In the SIR model, a node can be in one of the three states and thus the Markov chain has $N = 3^n$ vertices. Note that in the case of a complete graph, by symmetry arguments, the chain is exponentially smaller; one needs to keep track of the number of nodes in each state. A succinctly specified partially observable Markov decision process (POMDP) N can be obtained by augmenting CGDDS $\mathcal{G}$ with an appropriate model for observation $\mathcal{O}$ and decision-making $\mathcal{D}$.

Two basic problems in network epidemiology are, given $\mathcal{G}$, to characterize the structure of M, and finding appropriate upper and lower bounds for computing structural properties of M. It is now well accepted that the structure of the underlying contact graph G has a significant impact on the disease dynamics.[37] A fundamental question is to characterize the conditions under which there is a "large" outbreak (that is, when the number of infections is $\theta(n)$, 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 Erdos-Renyi[44] implies there is a large outbreak with $\theta(n)$ infections, if and only if $\beta > 1/n$.

The main technique is a branching process analysis.[16,37,44] We note the 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. It has been shown that such a threshold for disease spread exists in other well-structured graph classes, such as lattices and random regular graphs.[37,44] It is much more challenging to prove such statements in heterogeneous graphs with less symmetry. Pastor-Satorras et al.[39] use techniques from

Figure 3. The main steps in the first principles-based construction of synthetic populations and social contact networks.

Step 1 constructs a synthetic population by using various commercial and open source databases. Step 2 assigns daily activities to individuals within a household using activity and time-use surveys[5,21] as well as information available from social media. Step 3 constructs a *dynamic social bipartite visitation 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. (Image courtesy of Rachel Robinson.)



statistical mechanics to show that the threshold for epidemics propagation is 0 in scale-free network models, under mean-field assumptions, that is, no matter how small the probability of infection is, there would be a large outbreak. There are two settings for which rigorous results are known without the use of any mean-field assumptions. One is the Chung-Lu model,[12] 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 **w**. The other is classes of expander graphs.[1]

Computational Models

We outline here a general computational approach for networked epidemiology. The first practical and urban-scale application of this approach was described in Eubanks et al.[21] It consists of six broad steps (see "Computational Epidemiology over Networks"). Here, we discuss Steps 1–5; Step 6 will be detailed later. Please refer to the supplementary information[44] for details.

Steps 1–3 synthesize a realistic social contact network of the region under consideration (See Figure 3). Note that it is impossible to build synthetic urban-scale social contact networks solely by collecting field data, although such data can be incorporated into the synthetic population creation process. The networks so constructed are quite different structurally than those produced by simple random graph techniques.[5,22] Interesting similarities as well as differences among urban regions can be found in Barrett.[5] Recently, researchers have included other forms of data and information to extend the basic methodology described here. Examples include: using information from airline data to construct network-based representations of cities across the globe—each node corresponds to a city and the weight of each edge corresponds to the number of travelers that go between the two cities as measured by air transport data;[13] representation of smaller subnetworks (aka micro-networks), using either survey data or data collected using sensors,[31,41] and 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.[23,30] Note that the general approach is extensible to develop other kinds of affiliation networks, for example, individuals visiting a website or using a resource. Moreover, individual agents can be animals, devices, or digital objects. Finally, other attributes can be assigned to individuals so as to study other dynamic processes (assigning cellphones or demand for electricity). See supplementary information[44] for examples.

Step 4 is usually done in close coordination with biologists, epidemiologists, and statisticians. Computationally, this is naturally represented as a finite state probabilistic timed transition system; this is a finite state machine wherein certain transitions are timed and probabilistic. New methods in Bayesian inference as well as advances in genetics provide new opportunities for rapid inference.[20]

Step 5 involves the development of high-performance computing simulations to compute disease dynamics. This is a very active research area; the size of the networks and their unstructured and dynamic nature make for a challenging problem. Several researchers have developed scalable simulations of epidemic processes.[4,6,7,21,30,40]

A common approach to capture some of the individual decision-making in response to epidemics and public health advisories for vaccination is to study game theoretical formulations.

Control and Resource Optimization

Step 6 involves reasoning about POMDP N given $\mathcal{G}$, $\mathcal{O}$ and $\mathcal{D}$. Several interesting combinatorial formulations have been developed for vaccination and quarantining problems for epidemics on graphs, which we describe here briefly. Consider an SIR epidemic process defined on a graph G with probability p of transmission on any edge. We model the vaccination of a subset S of nodes in graph G by deleting them. Given a distribution $\mathbf{I}_0$ of the initial infections (so that Pr[v is infected initially] = $\mathbf{I}_0(v)$), and a budget B, the vaccination problem involves choosing a subset $S \subset V$ to vaccinate with $|S| \leq B$, so that the expected number of infections (due to the initial distribution $\mathbf{I}_0$) in $G[V - S]$ is minimized. The simplest setting for this problem is when $p = 1$, which might model a highly contagious disease. Even this problem is quite difficult to approximate, and bicriteria approximations have been developed for this problem, which choose a set S of size $(1+\epsilon)B$, and ensure the number of infections is at most $(1+1/\epsilon)$ times the optimal, for any $\epsilon > 0$.[22] The vaccination problem is much better understood in the SIS model, because of the characterization in terms of the spectral structure, as mentioned previously. The problem of designing interventions in this model can be reduced to controlling the eigenvalues.[8]

These results, though theoretically very interesting, are not realistic enough to be practical; we briefly discuss some of the issues that need to be incorporated in more realistic formulations. First, the general problem of interest is when $p \in (0, 1)$ and G is an arbitrary given graph. Second, in practice, interventions are a combination of vaccination and sequestration and social distancing—the sequestration of a subset $S \subseteq V$ can be modeled by the deletion of the edges in the cut $(S, \bar{S})$, with a corresponding cost for such deletions. Finally, interventions have a temporal dimension, since they are not implemented on the first day of the outbreak, but gradually, based on its progression. Further, partial outbreak information can be used to localize the current infections, and develop more efficient interventions. Finally, the contact graph and the disease model are only partially known and are dynamic.

A common approach to capture some of the individual decision-making in response to epidemics and public health advisories for vaccination is to study game theoretical formulations. We describe an early result by Aspnes et al.,[2] which has been studied quite extensively. This is formulated as a non-cooperative game, in which the strategy $x_v \in [0, 1]$ for a node v is the probability of getting vaccinated; here, we focus on pure strategies, in which $x_v \in \{0, 1\}$. For a given strategy vector $\mathbf{x}$, let G_x denote the graph obtained by deleting all nodes with $x_v = 1$. We assume an infection model of a highly contagious disease, in which all nodes reachable from an infected node get infected; further, we assume that the disease starts at a random initial node. Given a strategy vector $\mathbf{x}$, each user v has two kinds of costs: cost of getting vaccinated, $x_v C_v$, where C_v could be the economic or physical cost of the vaccine; and cost of infection, which is the probability that node v gets infected. This can be seen to be $|A(v)|^2/n$, where $A(v)$ denotes the connected component containing node v in the graph G_x, because of our infection model. It was shown this model always has a pure Nash equilibrium,[2] and the best social and Nash strategies can be approximated within $O(\log n)$ factor.

Behavioral adaption. The primary goal of an epidemiologist is to control the spread of infectious disease through the application of interventions guided by public policy. These interventions induce a behavioral change in individuals. At the same time individuals self-impose behavioral changes in response to their perception of how the disease evolves. Network-based epidemiology provides a natural representation to incorporate individual, collective, and institutional behavior (policies and responses) to control pandemics.

Verbal or conceptual behavioral models have played an important role in understanding behaviors and their relationship with diseases and maintaining a healthy life. Some of the earliest works in this direction are the Health Belief Model (HBM),[24] the Social Cognitive Theory (SCT),[3] and more recently the Social Ecological Model (SET). Verbal, social, and behavioral theories are useful in improving public health but are often informal. When using these models to develop formal computer models, these theories need to be "instantiated." This requires three components: algorithmic (mathematical) representation and computational resources as a basis of formalism, clarification of behaviors that occur at multiple scales: individual, group (for example, households, workplace), and organization (counties, companies), and expressiveness of behavioral representations from a computing standpoint. Obtaining data and conducting experiments to develop behavioral models is a challenging issue.[18,24] Recently, we developed a multitheory, multinetwork representation (see Figure 4) of individual, collective, and institutional behaviors to study the problem of anti-viral allocation during an epidemic (see supplementary information[44]).

Inference, Forecasting, and State Assessment

At the onset of an outbreak (for example, 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. 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 most 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,[38] because surveillance data on who is infected is limited and noisy (only a fraction of the infected people are detected); and 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. We briefly discuss a few important results but this largely remains an open topic.

A number of researchers[25,36,42] have studied these problems. An active area of research is based on Bayesian approaches, for example, using spatial scan statistics for detecting anomalous outbreaks;[36] these techniques rely on detecting spatial clusters from syndromic surveillance data, without using any specific properties about the epidemic process.

A rigorous result related to the inference problem is by Shah et al.,[43] who developed 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 order. They also developed 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 also been studied;[25] their results give a maximum likelihood estimator for a minimal network on which a given set of observed infections can occur.

Another important problem is to forecast the time course of an epidemic. The forecasting problem is complicated because behaviors, disease dynamics, and social networks coevolve. Researchers have studied various measures of effectiveness in this context, which include: peak value of an epidemic curve; time to reach the peak; total number of infections caused over a period (integral of the epidemic curve); the epidemic curve time series; and forecasting the onset of a pandemic. The forecasting models are then integrated with healthcare logistics to forecast required health care resources (for example, the number of beds, face masks, ventilators, and so on). The models are also integrated with pharmaceutical supply chains to develop a plan for production and distribution of vaccinations and anti-virals (for example, how many doses of vaccines to produce, availability, and distribution).[44]

Recent Advances

The role of computing goes beyond the topics we discussed so far. Recent advances in computing create exciting new opportunities for combining computational thinking and traditional epidemiology. Here, we discuss three salient topics:

Synthetic information and decision analytics. The epidemiological modeling tools provide detailed information on spatiotemporal disease dynamics. The data produced by these simulations as well as the data collected from various surveillance sources to initialize and calibrate these models poses challenging data analytics questions. Methods are required to analyze the outputs of the causal models to discov-

Computational Epidemiology Over Networks

Step 1. Construct a synthetic yet realistic population by integrating a variety of commercial and public sources.

Step 2. Build a set of detailed activity templates for households using time-use surveys and digital traces.[5,21]

Step 3. Construct a dynamic social bipartite visitation network, G_{PL}, which encodes the locations visited by each person.

Step 4. Develop models of within-host disease progression using detailed case-based data and serological samples to establish disease parameters.

Step 5. Develop high-performance computer simulations to study epidemic dynamics (exploring the Markov chain M).

Step 6. Develop multitheory multinetwork models of individual, collective, and organizational behaviors, formulating and evaluating the efficacy of various intervention strategies and methods for situation assessment and epidemic forecasting.

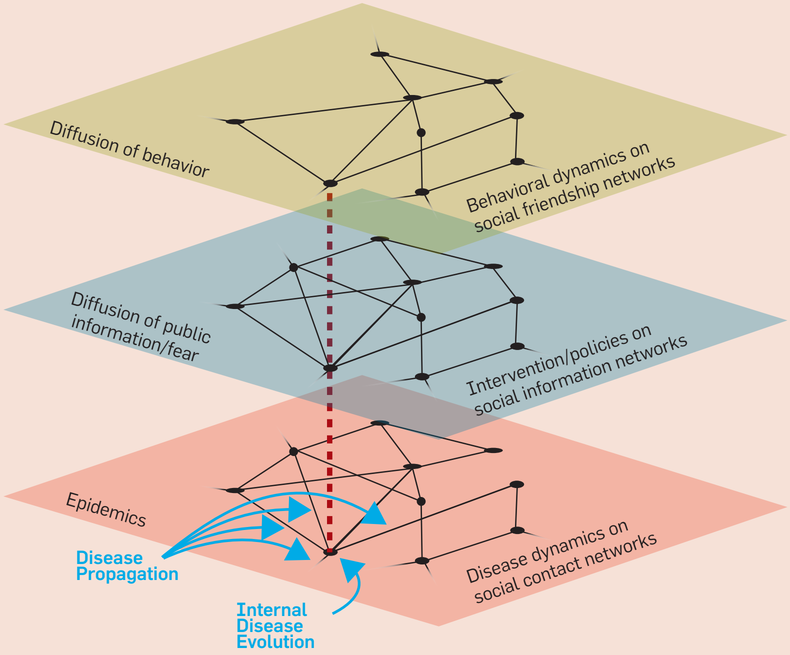
er new transmission chains, influential and critical nodes, optimal intervention pathways, among others.[2,21,22,25,36,43] Furthermore, these methods must be made accessible to epidemiologists and policymakers via a user-friendly environment that provides an easy way to set up experiments and analyze the results. Recently, a number of visual and data analytics methods and environments have been developed to support epidemiological research.[10,15,29] For example, a tool called the Interface to Synthetic Information Systems (ISIS) developed by our laboratory allows a user to set up detailed factorial experiments.[15,44] Using a simple interface to an underlying digital library, a user can choose from among many preconstructed instances: a social contact network; a within-host disease progression model; and a set of interventions. A key aspect of ISIS is its simplicity; we can train public health analysts to use the system in about three hours.[44]

Pervasive cyber-environment for computational epidemiology. Researchers have started developing a pervasive computing environment to support public health epidemiology.[10,15] Our group has also taken a step in this direction and developed a preliminary cyber-environment called Cyber Infrastructure for EPIdemics (CIEPI), which aims to make HPC resources seamless, invisible, and indispensable in routine analytical efforts. It is also designed to collect and process real-time data available via sensors and the Web. Indemics synthetic network construction methods and ISIS are integral components of CIEPI.

Big data and computational epidemiology. Recent advances in social media, computational turks, online games, online surveys, and digital traces all form the basis of potentially exciting new methods that can support surveillance, forecasting and tracking behavioral adaptation. For example, in the case of forecasting, the basic underlying hypothesis is that there is a strong correlation between incidence of diseases and the propensity of individuals to search for specific disease- related information or talk about it online.[11,19,42] Big data will play an increasingly important role in supporting computational epidemiology. Social media data, online micro-surveys and games, computationally mediated labor markets (for example, Mechanical Turks), improved bio-sensors, serological and genomic data will all lead to improved state assessment. These technologies pose new challenges, including: methods for eliciting truthful behavior; overcoming biases in data due to the demographics of participants; and translating behaviors from the virtual world to the real world.[42]

Figure 4. Multinetwork representations and coevolution of behavioral models.

The layers represent examples of different kinds of networks in which the nodes might be involved. The bottom layer is a social contact network formed by colocation constraints, on which diseases spread. The middle layer is an information network, on which information/fear spread. Finally, the top layer is a friendship network that spreads influence, for example, peer pressure. (Image courtesy of Henning Mortveit.)


Diffusion of behavior
Behavioral dynamics on social friendship networks
Diffusion of public information/fear
Intervention/policies on social information networks
Epidemics
Disease Propagation
Disease dynamics on social contact networks
Internal Disease Evolution

Policy Informatics

As discussed in Longini[26] and Lipsitch,[28] during the early stages of a pandemic, local, national, and global public health authorities need to make important decisions on allocating resources for tracking and controlling a potential pandemic beyond those required for routine disease monitoring. Allocation of these resources depends on two factors: estimating the risk of severity of the pathogen causing the infections, and the risk that the outbreak will reach a given region and cause widespread, serious illness. Unfortunately, it is extremely difficult to accurately estimate all of these risks, especially in the early stages of the pandemic. Hence, the role of surveillance as well as agile modeling and decision support environments to support analysis of counterfactual scenarios becomes all the more important. Appropriate response requires a delicate balance between slow and inadequate action on one hand and overreaction on the other hand. Additionally, a close collaboration with experts and decision makers who are often separated by institutional and governmental boundaries is important.

The computational tools we have built have been used to support a number of stakeholder-defined case studies. For example, as a part of the NIH-funded Models of Infectious Disease Agent Study (MIDAS)[35] network, we carried out a computational analysis of targeted layered containment strategy to control influenza pandemic. Results of the study were reviewed in a Letter Report by the Institute of Medicine.[32] We recently used CIEPI and associated tools to support near-real time decisions during the 2009 H1N1 pandemic. As H1N1 influenza continued to spread in the U.S., the Department of Health and Human Services teamed up with the Defense Threat Reduction Agency to place the ISIS tool in the hands of U.S. government analysts to provide day-to-day modeling results. Providing analysis inside the 24-hour decision cycle to this emerging crisis would not have been possible without the development of highly optimized modeling software as well as the Web-enabled interface. The analysts were able to perform course-of-action analyses to estimate the impact of closing schools and shutting down workplaces. Better situational awareness was enabled by calibrating the model to real-world data with different disease characteristics to estimate likely pathogen behaviors.

Summary

We outlined how recent advances in computing and information combined with computational thinking can be effectively employed to support public health epidemiology as it pertains to communicable diseases. Epidemiology of noncommunicable diseases (for example, obesity, diabetes) as well as environmental epidemiology (such as diseases caused by industrial toxic waste) are important but not discussed in this article. An important topic that we did not discuss is validation; see supplementary information[44] for details. Another topic not discussed is disease evolution and coevolving multistrain diseases. The ideas are also applicable in two related areas: zoonotic diseases and vector-borne diseases.[33,34]

We conclude by briefly discussing how concepts and ideas in epidemiology over the last 50 years have found applications in a large number of seemingly unrelated areas. The SIR model corresponds to the graph reliability problem, in which the goal is to determine the probability that a subset $S \subset V$ in a graph $G = (V, E)$ is connected, if each edge $e \in E$ fails independently with probability $p(e)$. The general ideas of epidemic diffusion have been used to design local and distributed algorithms for a variety of problems, including: routing, aggregation, broadcasting, and spatial gossip in communication networks (especially sensor networks),[17] synchronization, information sharing, and load balancing in parallel computing and replicated databases.[14] A broader class of reaction-diffusion models, which generalize SIR/SIS disease models, have been studied in diverse areas, including economics, sociology, viral marketing, political science, and social networks.[16]

Acknowledgments

The article represents the collective work of members of the Network Dynamics and Simulation Science Laboratory (NDSSL) and our external collaborators. We thank the reviewers for their feedback and suggestions. We thank Moshe Vardi for inviting us to submit the article and for his patience and willingness to work with us, without which the article would never have been completed.

The work is partially supported by NIH MIDAS, DTRA CNIMS, DOE ASCR, NSF ICES, NSF NetSe, NSF SDCI and IARPA programs.

References

1. Alon, N., Benjamini, I. and Stacey, A. Percolation on finite graphs and isoperimetric inequalities. *Annals of Probability* (2004).
2. Aspnes, J., Chang, K.L. and Yampolskiy, A. Inoculation strategies for victims of viruses and the sum-of-squares partition problem. *J. Comput. Syst. Sci.* (2006).
3. Bandura, A. *Social Foundations of Thought and Action: A Social Cognitive Theory.* Prentice Hall. 1986.
4. Barrett, C. et al. Episimdemics: An efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *Proc. of the ACM/IEEE Conference on High Performance Computing* (2008).
5. Barrett, C. et al. Generation and analysis of large synthetic social contact networks. In *Proc. Winter Simulation Conference* (2009), 1003–1014.
6. Bisset, K. et al. Indemics: An interactive data intensive framework for high performance epidemic simulation. In *Proc. Int. Conf. Supercomputing* (2010).
7. Bisset, K. et al. Interaction-based HPC modeling of social, biological, and economic contagions over large networks. In *Proc. of Winter Simulation Conference* (2011), 2933–2947.
8. Borgs, C., Chayes, J.T., Ganesh, A. and Saberi, A. How to distribute antidotes to control epidemics. *Random Structures and Algorithms* (2010).
9. Brauer, F., van den Driessche, P. and Wu, J. editors. Mathematical Epidemiology. *Lecture Notes in Mathematics.* Springer Verlag, 1945.
10. Broeck, W. et al. The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC Infectious Diseases 11*, 1:37 (2011).
11. Brownstein, J., Freifeld, C. and Madoff, L. Digital disease detection—harnessing the Web for public health surveillance. *N. England J. Med.* (2009), 2153–2157.
12. Chung, F. and Lu, L. Connected components in random graphs with given degree sequences. *Annals of Combinatorics 6* (2002), 125–145.
13. Colizza, V., Barrat, A., Barthelemy, M. and Vespignani, A. The role of the airline transportation network in the prediction and predictability of global epidemics. *PNAS 103*, 7 (2006), 2015–2020.
14. Demers, A. et al. Epidemic algorithms for replicated database maintenance. In *Proceedings for ACM PODC* (1987), 1–12.
15. Deodhar, S. et al. Enhancing user-productivity and capability through integration of distinct software in epidemiological systems. *IHI* (2012), 171–180.
16. Easley, D. and Kleinberg, J. *Networks, Crowds and Markets: Reasoning About A Highly Connected World.* Cambridge University Press, New York, 2010.
17. Epidemic Routing Bibliography. http://roland.grc.nasa.gov/~weddy/biblio/epidemic/.
18. Cowling, B. et al. Community psychological and behavioral responses through the first wave of the 2009 influenza a (H1N1) pandemic in Hong Kong. *J. Infect. Diseases 202* (2010), 867–877.
19. Ginsberg, J. et al. Detecting influenza epidemics using search engine query data. *Nature* (2008), 1012–1014.
20. Cauchemez, S. et al. Household transmission of 2009 pandemic influenza a (H1N1) virus in the United States. *New England J. Medicine 27* (2009), 2619–2627.
21. Eubank, S. et al. Modeling disease outbreaks in realistic urban social networks. *Nature 429* (2004), 180–184.
22. Eubank, S. et al. Structure of social networks and their impact on epidemics. *DIMACS Discrete Methods in Epidemiology* (2006), 179–185.
23. Ferguson, N.M. et al. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature 437* (2005), 209–214.
24. Funk, S., Salathé, M. and Jansen, V. Modeling the influence of human behaviour on the spread of infectious diseases: A review. *J. R. Soc. Interface 7* (2010), 1247–1256.
25. Gomez-Rodriguez, M., Leskovec, J. and Krause, A. Inferring networks of diffusion and influence. In *Proc. 16th ACM KDD*, 2010.
26. Van Kerkhove, M. and Ferguson, N. Epidemic and intervention modeling—A scientific rationale for policy decisions: Lessons from the 2009 influenza pandemic. *Bull World Health Organ.* (2012), 306–310.
27. Lipsitch, M. and et al. Managing and reducing uncertainty in an emerging influenza pandemic. *New England Journal of Medicine 361*, 2 (2009), 112–115.
28. Lipsitch, M. et al. Improving the evidence base for decision making during a pandemic: The example of 2009 influenza-A H1N1. *Biosecur Bioterror.* (2011).
29. Livnat, Y., Rhyne, T. and Samore, M. Epinome: A visual-analytics workbench for epidemiology data. *IEEE Comp. Graphics & App. 32*, 2 (2012), 89–95.
30. Longini. I.M. et al. Containing pandemic influenza at the source. *Science 309* (2005), 1083–1087.
31. Mossong, J. et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 5, 3 (2008) e74.
32. National Academies. Modeling Community Containment for Pandemic Influenza: A Letter Report (2006).
33. National Academies. The Causes and Impacts of Neglected Tropical and Zoonotic Diseases: Opportunities for Integrated Intervention Strategies (2011).
34. National Academies. Vector-Borne diseases: Understanding the Environmental, Human Health, and Ecological Connections (2011).
35. National Institutes of Health, 2009; http://www.nigms.nih.gov/Initiatives/MIDAS/.
36. Neil, D., Moore, A. and Cooper, G. A Bayesian spatial scan statistic. NIPS (2005).
37. Newman, M. The structure and function of complex networks. *SIAM Review 45* (2003).
38. Nishiura, H. et al. Did modeling overestimate the transmission potential of pandemic (H1N1-2009)? Sample size estimation for post-epidemic seroepidemiological studies. PLoS ONE 3 (2011).
39. Pastor-Satorras, R. and Vespignani, A. Epidemic spreading in scale-free networks. *Phys. Rev. Lett. 86* (Apr. 2001). 3200–3203.
40. Perumalla, K. and Seal, S. Discrete event modeling and massively parallel execution of epidemic outbreak phenomena. *Simulation*, 2011.
41. Salathé, M. et al. A high-resolution human contact network for infectious disease transmission. *PNAS 107*, 51 (2010), 22020–22025.
42. Salathé, M. et al. Digital epidemiology. *PLoS Computational Biology 8*, 7 (2012).
43. Shah, D. and Zaman, T. Rumors in a network: Who is the culprit? *ACM SIGMETRICS* (2010).
44. Supplementary information; http://ndssl.vbi.vt.edu/supplementary-info/vskumar/cacm2012/.

Madhav Marathe (mmarathe@vbi.vt.edu) is a professor in the Department of Computer Science and Network Dynamics and Simulation Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA.

Anil Kumar S. Vullikanti (akumar@vbi.vt.edu) is an associate professor in the Department of Computer Science and Network Dynamics and Simulation Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA.