Computational Epidemiology in a Connected World

Andrea Apolloni, V.S. Anil Kumar, Madhav V. Marathe, and Samarth Swarup, *Virginia Tech*

New technologies help epidemiologists model the socioeconomic context of disease outbreaks.

In 1854, a physician named John Snow helped end a deadly cholera outbreak in London's Soho district. The bacterium that causes cholera had not yet been identified, and very little was known about how the disease spreads.

Snow gathered information about the local infrastructure, people's sanitary and social habits, and demographic data such as residents' profession, age, and socioeconomic status and determined that the most likely cause of the outbreak was water from a particular pump. He then persuaded the authorities to remove the pump's handle, preventing further spread of the disease.

Snow's genius lay in his ability to combine all the social, economic, geographical, and biological data available—though often limited and anecdotal—to infer the mode of transmission as well as the source of the outbreak. His achievement is widely considered one of the founding events of modern epidemiology.

Snow's essential insight, that stopping an epidemic requires understanding its socioeconomic context, is even more relevant in today's world. The medical and public health communities have made tremendous strides to detect, respond to, and control epidemics—the worldwide coordinated efforts that contained the 2002-2003 SARS virus are a testimony to this. Nevertheless, pandemics such as the recent H1N1 influenza will continue to occur, exacerbated by global trends such as increasing urbanization, travel, and immunocompromise.

Epidemics place a huge cost upon society. The 1918 flu pandemic caused some 50 million deaths worldwide, and it is estimated that a similar pandemic today would result in 150 million deaths and cost $4.4 trillion.

Epidemiologists and computer scientists are developing new data-driven, high-performance-computing-powered inference engines to model the socioeconomic context and strategies necessary to counter disease outbreaks.

EPIDEMICS AS COMPLEX SYSTEMS

Infectious diseases often spread throughout social networks as those who are infectious come in contact with susceptible individuals. To predict an epidemic's course, researchers must therefore track the health status as well as the movements and interactions of people as they carry out their daily activities.

Social contact networks, however, tend to be highly dynamic. Several aspects of this change are endogenous—people's daily schedules are affected by, among other things, the epidemic itself. If a person is ill, he might decide to stay home from work; conversely, if he fears contracting the disease at work, he may decide to stay home too. These behavioral adaptations alter social contact networks and in turn impact the epidemic's progress.

Pharmaceutical interventions, such as antiviral drugs and vaccines, and public policy responses, such as school closures and social distancing, likewise affect social contact networks. Policies must maintain a delicate balance between disease control on one hand and normal societal functioning on the other. Overreactive policies reduce public confidence and compliance, while delayed actions increase the spread of the disease. Thus, an epidemic is not simply a diffusion over a network,

0018-9162/09/$26.00 © 2009 IEEE Published by the IEEE Computer Society

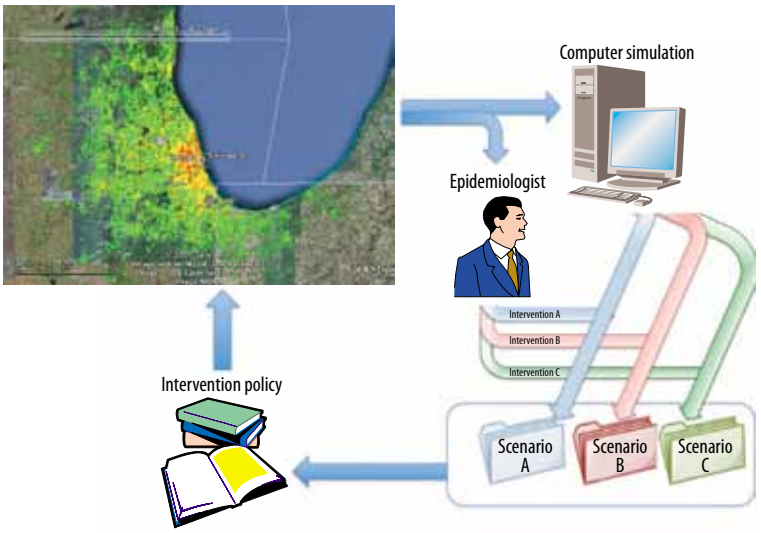

Figure 1. Measure-project-intervene cycle. Epidemiologists use data gathered from surveillance to seed computer simulations of an epidemic's progression. They then use these simulations to evaluate a range of possible interventions and establish intervention priorities. Data from intervention outcomes in the real world is then gathered for the next cycle.

but a coevolving system of multiple networks, dynamic processes (spread of disease, fear, and so on), and individual behavioral adaptation.

Once the number of infected individuals crosses a certain threshold that depends on complex sociotechnical and disease-specific variables, it is virtually certain an epidemic will occur. Effective epidemiological control therefore requires a rapid but pragmatic response, given limited and highly uncertain information.

For example, the first H1N1 case was diagnosed in Mexico on 17 March 2009, though there is evidence that the so-called "swine flu" virus had been spreading several months before that. By the end of March, the disease had already appeared in California. In April, world public health officials recognized the outbreak's severity, though data about transmissibility and mortality rates was still unavailable.

Mexico City authorities began closing down schools and public places, as did officials in Texas. Despite these efforts, H1N1 spread rapidly. Epidemiologists issued their first evaluation of the outbreak's seriousness in early May, and the following month the World Health Organization declared a global pandemic. Thus, by the time researchers had hard data about H1N1, people throughout large parts of the world were already infected.

Most pandemic responses thus focus on mitigation rather than prevention. However, school closings and other such efforts can often have unforeseen consequences. For example, when schools close, parents might simply send their children to daycare instead, which does not serve the intended goal of reducing contact between children.

This combination of complex evolving social interaction, limited and delayed information, and unforeseen consequences of interventions makes computational epidemiology among the hardest problems in science and policy.

TRADITIONAL APPROACH

Mathematical epidemiology has traditionally relied on rate-based differential-equation models. In this approach, researchers partition a population into subgroups based on various criteria, such as demographic characteristics and disease states, and use the models to describe disease dynamics across these groups.

One of the earliest analyses was by R. Ross, who studied the spread of malaria in the late 19th century. W.O. Kermack and A.G. McKendrick further developed this technique in the 1920s and 1930s to investigate short-term diseases like measles and influenza. They showed that disease dynamics are characterized by a parameter R_0, the basic reproduction number—defined as the number of secondary infections caused by a single infective into a wholly susceptible population. If $R_0 < 1$, the infection will die out; if $R_0 > 1$, an epidemic will occur.

This approach has been tremendously successful in informing public health policy. Nevertheless, a potential weakness is its inability to capture the complexity of human interaction and behavior.

MEASURE-PROJECT-INTERVENE CYCLE

Effective epidemiology is not just about prediction, but also about anticipation and adaptation. The difference is akin to that between golf and basketball. A golf course is essentially static, and golfers must carefully evaluate the prevailing conditions such as wind speed and direction, as well as the topography, before hitting the ball. Importantly, the act of hitting the ball doesn't change the course conditions. In basketball, on the other hand, each player is constantly in motion, and decisions about passing, shooting, and other movements depend on all the players' current and anticipated future actions.

Epidemic planning and response similarly does not involve measuring the disease conditions and then acting once. Rather, as Figure 1 shows, there is a continuous measure-project-intervene cycle not unlike the sense-evaluate-act cycle of a cognitive agent.

The measuring step involves gathering surveillance data from healthcare agencies. Usually, only partial information is available from such agencies, leading to Bayesian inference problems in determining the epidemic's source and current state.

Unfortunately, the systems of interest have extremely large state spaces—a social contact network modeling a moderate-size city with a million inhabitants must efficiently process a state space with $2^{1,000,000}$ states. This is clearly infeasible and motivates further research on graphical models for Bayesian inference that can take the problem semantics into account to reduce the effective state space.

An alternative approach uses agent-based models as part of the predictive filter for situation assessment. The projection step involves doing hundreds of simulations of possible intervention scenarios to find the one with the highest likelihood of steering the system in the right direction. The interventions are adaptive as a result of constant behavioral changes and thus are naturally represented as Markov decision problems.

Human experts design the intervention scenarios, choose the initial conditions, and incorporate the noisy, delayed, and incomplete surveillance data. The computational system provides a quantitative assessment of each strategy's strengths and weaknesses. Policymakers use this assessment in conjunction with various complex socioeconomic constraints to select and implement the final strategy. The measure-project-intervene cycle then continues.

Recent public health policies and their modifications pertaining to school closures and vaccine allocations in the context of the H1N1 outbreak are good examples of this model-based reasoning process at work.

INTERACTION-BASED APPROACH

The measure-project-intervene cycle motivates an interaction-based approach that involves accurate modeling of social interaction networks and disease dynamics (C.L. Barrett, S. Eubank, and M. Marathe, "An Interaction-Based Approach to Computational Epidemiology," *Proc. 23rd Nat'l Conf. Artificial Intelligence*, vol. 3, AAAI Press, 2008, pp. 1590-1593). This approach combines endogenous representations of individuals with explicit interactions among them to generate and capture a disease's spread across the social interaction network.

The biggest strengths of the synthetic information environment approach are its scalability and its extensibility.

The interaction-based approach goes beyond traditional mathematical modeling techniques, which assume homogeneous interactions within each population segment. It also raises new technical difficulties: It is impossible to obtain accurate, detailed, time-varying, urban-scale social contact networks by simple measurement. Nevertheless, recent advances in computing technology, machine learning, data mining, and network science make it possible to develop new approaches for producing reasonable estimates of such networks.

We have developed one such computational approach that uses *synthetic information environments*. An SIE consists of

- a statistical model of the population of interest, known as a *synthetic population*;
- an activity-based model of the social contact network;
- models of disease progression; and
- models for representing and evaluating interventions, public policies, and individual behavioral adaptations.

The biggest strengths of the SIE approach are its scalability and its extensibility. An epidemiologist using the system can easily design a new intervention and run the corresponding simulation for a large urban area like Los Angeles in minutes. From data analysis she can find critical pathways as well as assess the indirect effects—for example, the economic impact—of certain policies.

SIMDEMICS

The Network Dynamics and Simulation Science Laboratory (NDSSL) at Virginia Tech's Virginia Bioinformatics Institute has developed Simdemics, an integrated modeling environment that aids state, local, and federal public health officials in pandemic planning, response, and control. Simdemics' computer models embody all four SIE components. NDSSL team members are also developing Isis, a service-oriented computing environment that lets users seamlessly access Simdemics using today's Web technology.

Researchers validate and verify complex software environments such as Simdemics using *composite validity* and *adequacy*. Standard validation techniques based on matching historical data are usually not meaningful in this context.

We have used Simdemics to explore numerous important research questions. For example, a recent study of the socioeconomic impact of various intervention strategies aimed at controlling an influenza-like illness showed that a combination of school closures, individual context-based behavioral adaptation, and targeted antiviral distribution can reduce the disease's overall impact by 87 percent and income loss by 82 percent as compared to the base case (C.L. Barrett

et al., "Estimating the Impact of Public and Private Strategies for Controlling an Epidemic: A Multi-Agent Approach," *Proc. 21st Innovative Applications of Artificial Intelligence Conf.*, AAAI Press, 2009, pp. 34-39).

A century and a half after Snow helped found the science of epidemiology, his ideas apply even more broadly. Today policymakers must minimize the economic and social impact of an epidemic as well as the disease itself. Epidemics are complex systems, and as such their behavior is often unintuitive. Planning for and controlling them requires considerable experience and a willingness to compromise, necessitating increasingly sophisticated models and computational support for years to come.

Andrea Apolloni *is a postdoctoral research associate at the Network Dynamics and Simulation Science Laboratory (NDSSL), Virginia Bioinformatics Institute, Virginia Tech. Contact him at apolloni@vbi.vt.edu.*

V.S. Anil Kumar *is an assistant professor in the Department of Computer Science and a senior research associate at the Virginia Bioinformatics Institute, Virginia Tech. Contact him at akumar@vbi.vt.edu.*

Madhav V. Marathe *is a professor of computer science and deputy director of NDSSL, Virginia Bioinformatics Institute, Virginia Tech. Contact him at mmarathe@vbi.vt.edu.*

Samarth Swarup *is a postdoctoral research associate at NDSSL, Virginia Bioinformatics Institute, Virginia Tech. Contact him at swarup@vbi.vt.edu.*

Editor: Naren Ramakrishnan, Dept. of Computer Science, Virginia Tech, Blacksburg, VA; naren@cs.vt.edu