



Recent Advances in Computational Epidemiology

Madhav V. Marathe and Naren Ramakrishnan, *Virginia Tech*

Epidemiology is witnessing a rapid infusion of new techniques from computer science, especially machine learning. To understand these new developments, it's helpful to contrast them against traditional approaches to studying disease progression, such as the mean field approach that uses rate-based differential equation models. In this approach, we partition the human population (although technically other species are also studied in epidemiology) into subgroups based on various criteria (for example, demographic characteristics and disease states), and use differential equation models to describe the disease dynamics across these groups. Some models characterize disease dynamics by a parameter, R_0 , the basic reproduction number.¹ R_0 is defined as the number of secondary infections caused by a single infective individual into a wholly susceptible population. R_0 determines whether an epidemic can occur: if $R_0 < 1$ the epidemic will die out, while 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 interactions and behaviors.

Effective planning and response in the event of epidemics isn't about just prediction, but anticipation and adaptation. The typical workflow of a public health analyst involves the measure-project-analyze-intervene cycle. In this method, diverse data is collected via surveys, social media, sensors, and policy documents, which are then analyzed to yield contextual situational representations. Dynamic models in the form of computer simulations are then used to interpolate as well as extrapolate from the data. Simulations are used to evaluate various what-if scenarios (or counterfactual experiments). Policy analysts use this information

to make specific policy decisions, potentially leading to changes in epidemic dynamics. The measure-project-analyze-intervene cycle motivates an interaction-based approach for developing informatics platforms. Here, we aim to accurately model the social interactions that form the basis of disease transmission. The approach uses endogenous representations of individuals together with explicit interactions between these agents to generate and capture the disease spread across the social interaction network.

However, this approach is fraught with new technical difficulties. It's impossible to obtain an accurate, detailed, time-varying, urban-scale, human social-contact network by simple measurements. Nevertheless, recent advances in machine learning, data mining, and network science make it possible to develop new approaches for producing reasonable estimates of such networks. We've developed one such computational approach, the synthetic information environments (SIEs) approach.

Synthetic Information Environments

An SIE consists of four components:

- statistical model of the population of interest, which we refer to as a *synthetic population*;
- activity-based model of the social-contact network;
- disease-progression models; and
- models for representing and evaluating interventions, public policies, and individual behavioral adaptations.²

First, we generate a synthetic population by integrating census data with other demographic and geographic data to create a population of individual agents. Synthetic populations are statistically

identical to the data sources that are used to construct them, but preserve individual privacy and maintain anonymity. Second, we generate a detailed minute-by-minute schedule for each individual in the synthetic population, using time-use surveys combined with machine learning techniques, such as classification and regression trees (CART). We geolocate activities using business survey information and, by employing a gravity model, associate each individual with particular activity locations over the course of the day. The availability of modern datasets collected via phone call logs and social media sites such as Foursquare provide new opportunities to refine the methodology and improve the assignment quality.

We can now construct a time-varying, spatially-explicit, person-location network using the synthetic data. The synthesis of such networks is an ongoing research theme in computational social science and is sometimes referred to as *generative social science*.³ Recently, researchers have explored other methods to synthesize smaller social contact networks using smartphones, RFID tags, and other digital devices combined with social media; examples include synthesis of social contact networks among high school and college students. These methods provide valuable data sources to create smaller subnetworks useful for validation purposes.⁴⁻⁷

In the third step, we endow each individual with a within-host disease model represented using probabilistic timed transition systems (PTTS). We can incorporate within the framework individual-level demographic variations (immunity, age, and so on). Individual PTTS are coupled via the social contact network described earlier. We use high-performance computer simulations to understand the spread of the contagion over the network of PTTS.

The final step involves representing and analyzing public policies, individual behavioral adaptations, and the efficacy of various intervention strategies. A key concept here is that of implementable policies and interventions—that is, policies that are realizable in the real world. For example, an optimal vaccination policy based on computational models might specify a set of k -individuals who are super spreaders and hence should be vaccinated. But in the real world, it's not easy to identify these individuals explicitly. We use data mining and machine learning techniques to identify surrogates (that is, combinations of demographic and social attributes) that can redescribe the super spreader property.

The biggest strengths of the SIE approach are its scalability and extensibility. An epidemiologist using the system can easily design a new intervention and carry out an appropriate computer experiment for a large urban area like Los Angeles in minutes to uncover critical individuals and pathways and evaluate the indirect effects (for example, the economic impact) of certain policies.

Simdemics is an integrated modeling environment that embodies the SIE approach to aid local, state, and federal public health officials in pandemic planning, response, and control.⁸ As an example, in other work we used Simdemics to estimate the social and economic impact of the various public and private intervention strategies aimed at controlling influenza-like illnesses.⁹ We developed a synthetic social contact network for the New River Valley area of Virginia. We evaluated a range of realistic, individual behavioral strategies as well as public policies to control a flu-like epidemic. The study showed that a combination of school closures, individual context-based behavioral adaptation,

and targeted antiviral medication distribution can reduce the number of infections by 87 percent and income loss by 82 percent as compared to the base case with no intervention.

Big Data and Real-Time Epidemiology

Real-time epidemiology, a rapidly developing area within public-health epidemiology, seeks to support policy makers in near real-time as an epidemic is unfolding.¹⁰ A natural use of real-time epidemiology is in disease surveillance, that is, monitoring the space-time progression of disease. Traditional tools for surveillance include sentinel clinics and serological sampling.¹¹ Recently, researchers have used social-media data to obtain disease outbreak and progression information, an excellent example of how computational advances are changing public health epidemiology.^{12,13}

Perhaps the most celebrated example of social media surveillance is Google Flu Trends (www.google.org/flutrends) that uses search engine queries as an indicator of health-seeking behavior, and thus an indicator of disease (flu) activity among a population.¹⁴ Not long after Google Flu Trends was introduced, techniques for nowcasting flu rates using Twitter became prominent.¹⁵ Researchers have paid careful attention to content modeling of tweets. For instance, Alex Lamb and colleagues have developed methods to separate tweets that report actual flu infections from others that exhibit mere awareness and concern about the flu.¹⁶ Broader uses of Twitter for syndromic surveillance—in particular for capturing spatiotemporal distributions of symptoms and medications—have also been explored.¹⁷ In general, social media is a fertile resource for exploring many epidemiological questions, for example, sentiment propagation about vaccinations.¹⁸

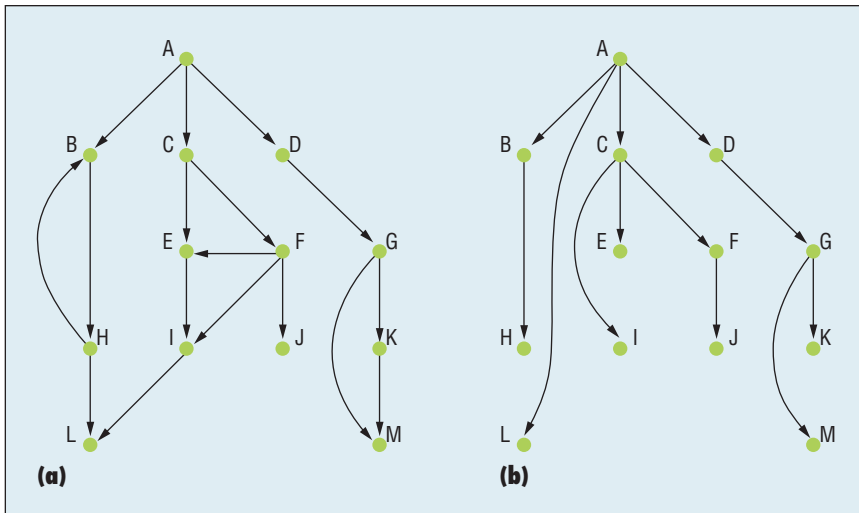


Figure 1. A social contact network where nodes represent people. (a) In the example graph, we see that all paths from node A to H must pass through B. Hence, B dominates H. (b) Uncovering dominance of nodes lets us create a dominator tree.

The previously discussed methods focus on gross estimation of disease activity over a region. In line with our earlier discussion about synthetic populations, researchers have also explored unraveling patterns of online communication from Twitter with a view to uncovering social interactions. Adam Sadilek and his colleagues use geolocation and machine-learning methods to estimate physical interactions between healthy and sick individuals and, in turn, estimate the likelihood of the healthy individual getting infected at some point in the future.¹⁹

More recent research has focused on identifying social network sensors, that is, identifying a subset of individuals whose infection states can be monitored to serve as an early indicator of an emerging epidemic. Nicholas Christakis and James Fowler²⁰ propose a design of social network sensors for monitoring flu on the basis of the friendship paradox: your friends have more friends than you do. Alternatively it can be said that a friend of a random person has more friends than the random person. Christakis and Fowler use the set of friends nominated by randomly chosen people as a sensor set. After a field study on randomly selected

students at Harvard during the flu season in 2009, they found that the peak of the daily incidence curve in the sensor set occurs 3.2 days earlier than that of a random set of students.

In other research, we formalized the idea of social network sensors using the notion of graph dominators.^{21,22} In a given graph, a node x is said to dominate a node y if all paths from a designated start node to y must go through x . In our case, the start node indicates the source of the infection or disease. In Figure 1a, which describes a social contact network with nodes as people, all paths from node A (the designated start node) to H must pass through B; therefore B dominates H. Note that a person can be dominated by many other people. For instance, both C and F dominate J, and C dominates F. To simplify such transitive situations, we say that node x is the unique immediate dominator of y if x dominates y and there doesn't exist a node z such that x dominates z , and z dominates y . This enables us to uncover an underlying tree of dominator relationships, as shown in Figure 1b, with a much smaller number of edges than the original graph.

If we were to reconstruct the social contact network, therefore, we

can readily compute the dominator tree and capture critical junctures in epidemic transmission. Using city-scale datasets generated by extensive microscopic epidemiological simulations involving millions of individuals, we've shown how the notion of dominators can provide up to 10 days more lead time compared to the friend-of-friends approach (see Figure 2). Most importantly, in other research, we developed surrogates and proxies for use as social-network sensors without requiring intrusive knowledge of people and their relationships.²¹ For instance, we can identify demographic properties that best redescribe the dominator relationship, and use these properties to help form the sensor set in practice.

Resource Allocation, Behavior Modeling, and Inference

Computational models and machine learning are important for broader policy questions in epidemiology as well. When applying these techniques in practice, researchers face the usual challenges: noisy and insufficient data, scarce resources, multiple objective functions, and short decision-making time.

Resource optimization problems arise in epidemiology when scarce public health resources need to be expended to respond to epidemic outbreaks. Examples of such problems include: allocation of vaccines and antivirals; availability of medical equipment such as facemasks, hospital beds, and ventilators; staffing problems at hospitals; and allocation of pharmaceuticals. The objective functions are complex, including economic costs, health costs, and social disruptions. Moreover, the objectives are usually conflicting, thus making the decision-making process harder.

Inference problems in epidemics arise from the need to understand the spatiotemporal characteristics of an epidemic, especially at the start of the epidemic. Examples include inferring the index case, disease properties, social contact network, and transmission tree.

A prototypical and important problem is vaccine allocation for controlling influenza outbreaks. Even the basic problem is computationally challenging. The issue is complicated by the

fact that various logistical complications cause vaccines to become available in batches. Moreover, just like in the social-network sensors problem, it's important to develop an implementable strategy for assigning vaccines. Classical work has focused either on optimal strategies that aren't implementable or on allocating vaccines to predefined groups. In other work, we combine data-mining techniques and dynamical properties of networks to design a near-optimal vaccination strategy that compares well with known strategies.²³

It's important to note that the application of interventions, guided by public policy, will in turn induce behavioral changes in individuals. A computational representation theory of behaviors as it pertains to epidemiology thus needs to be developed. Health scientists have developed verbal or conceptual behavioral models to understand the role of behaviors in public health.^{24,25} But these models are typically informal and it's quite demanding to identify the data necessary to instantiate in silico behavioral models. Recent advances in social media, crowdsourcing (for example,

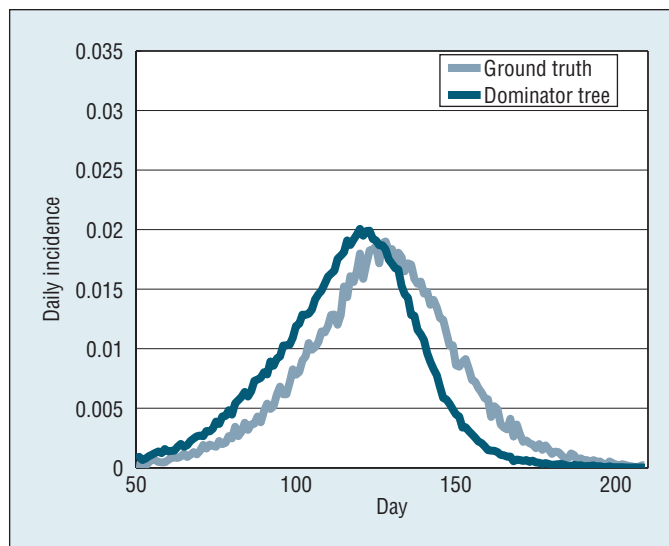


Figure 2. Monitoring an epidemic using a social network sensor on the basis of a dominator heuristic enables earlier detection—that is, the peak in the sensor curve occurs ahead of the peak in the general population.

Amazon's Mechanical Turk at www.mturk.com), online games, online surveys, and digital traces all form the basis of potentially exciting methods to make progress in this direction.²⁶ We've developed a computational modeling environment wherein complex behaviors and interventions can be represented and analyzed.²⁷ Figure 3 presents our system's interface that enables the analyst to set up complex statistical experiments (interventions) and analyze their effects on the underlying population. The experiments are then executed using a high-performance computing simulation, and the results are summarized and presented to the user.

As a case study, we've explored an important policy problem in epidemiology: Is there an optimal strategy to distribute a limited supply of antiviral doses between the public stockpile administered through hospitals and private stockpiles distributed through a market mechanism? In modeling this problem, we considered a number of measures of effectiveness, including number of people infected, peak number of infections, cost of recovery, and equitable allocation. We

were broadly interested in understanding how disease dynamics, individual behavior, network structure, and antiviral demand coevolve. We developed and instantiated several behavioral models based on published literature and data. These models spanned individual behaviors (for example, reporting of symptoms by infected persons), family behaviors (such as purchasing behaviors and isolation precautions), and organizational behaviors (including behavior of

markets as well as entities such as hospitals). Our other research provides more details.^{27,28}

Key findings based on our experiments include the following: market-based distribution is inherently inequitable; prevalence of elastic demand leads to inequitable distribution (due to price increase), providing ways to evaluate government investment; there is an optimal allocation strategy of antivirals between public and private stockpiles; and natural behavior adaptations in conjunction with well-established logistics (markets and public distribution) reduce and delay the peak infection rate.

The use of machine learning and reasoning methods in support of computational epidemiology is a rich area with many significant research challenges. Key areas for future research include the following:

- *New methods and data sources for extending synthetic populations.* This is a relatively understudied problem, and formal

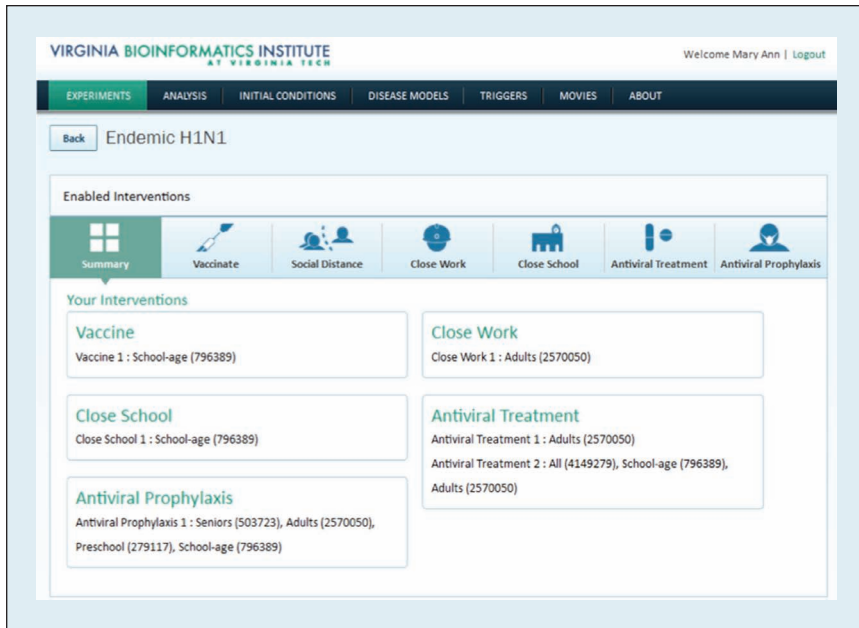


Figure 3. Isis is a Web-based decision-support environment that allows public health epidemiologists to analyze various counter-factual scenarios related to epidemic planning. The image shows the system's interface.

Acknowledgments

This work has been partially supported by National Science Foundation HSD grant SES-0729441, NSF PetaApps grant OCI-0904844, 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 number 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

1. W.O. Kermack and A.G. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," *Proc. Royal Soc. London A*, vol. 115, no. 772, 1927, pp. 700–721.
2. C.L. Barrett, S. Eubank, and M.V. Marathe, "An Interaction Based Approach to Computational Epidemics," *Proc. Ann. Conf. AAAI, AAAI*, 2008, pp. 1590–1593.
3. J. Epstein, *Generative Social Science: Studies in Agent-Based Computational Modeling*, Princeton Univ. Press, 2005.
4. L. Glass and R. Glass, "Social Contact Networks for the Spread of Pandemic Influenza in Children and Teenagers," *BMC Public Health*, vol. 8, no. 1, 2008, p. 61; doi:10.1186/1471-2458-8-61.
5. M. Salathé et al., "A High-Resolution Human Contact Network for Infectious Disease Transmission," *Proc. Nat'l Academy Sciences*, vol. 107, no. 51, 2010, pp. 22020–22025.
6. J. Stehle et al., "Simulation of an SEIR Infectious Disease Model on the Dynamic Contact Network of Conference Attendees," *BMC Medicine*, vol. 9,

characterization of the difficulty of the problem, as well as efficient and effective algorithm development, needs to be undertaken.

- *Integrating model-driven methods with data mining approaches.* We have hinted at some possibilities here, but more opportunities abound—for example, using a combination of approaches to design quarantine policies from field data, behavioral models, and a theory-driven statement of epidemiological objectives.
- *Social-network sensors.* Can we develop new methods and surrogates for identifying sensor populations from both massive passive data (Twitter) and for use in clinics and hospitals?
- *Fine-grained modeling of social-media datasets.* As techniques for content modeling and text mining become increasingly sophisticated, we believe there will be a greater carryover of such methods to syndromic surveillance with real-time epidemiological applications.


- *Active data collection, leading to coevolving policy, simulation, and mining.* There's increasing interest in conducting mobile phone surveys and integrating such survey data with more passively gathered information. Active data can help fill in information gaps from traditional data mining of passive datasets. For instance, a survey of disease symptoms in a targeted region combined with mining of tweets can give lead time advantages in detecting an emerging epidemic.

Tackling these problems will require a multidisciplinary approach and a close collaboration between computer scientists, statisticians, public health experts, and policy analysts. Finally, although we restrict our discussion to infectious diseases in humans, researchers can also study zoonotic diseases and many chronic diseases such as obesity and diabetes within this framework. We look forward to seeing future developments in this diverse and important field.

- no. 1, 2011, p. 87; doi:10.1186/1741-7015-9-87.
7. A. Madan et al., "Social Sensing for Epidemiological Behavior Change," *Proc. 12th ACM Int'l Conf. Ubiquitous Computing*, ACM, 2010, pp. 291–300.
 8. C. Barrett et al., "An Integrated Modeling Environment to Study the Co-Evolution of Networks, Individual Behavior, and Epidemics," *AI Magazine*, vol. 31, no. 1, 2010, pp. 75–87.
 9. C. Barrett et al., "Economic and Social Impact of Influenza Mitigation Strategies by Demographic Class," *Epidemics J.*, vol. 3, no. 1, 2011, pp. 19–31.
 10. H.V. Fineberg and M.E. Wilson, "Epidemic Science in Real Time," *Science*, vol. 324, no. 5930, 2009, p. 987; doi:10.1126/science.1176297.
 11. J. Brownstein, C. Freifeld, and L. Madoff, "Digital Disease Detection—Harnessing the Web for Public Health Surveillance," *New England J. Medicine*, vol. 360, no. 21, 2009, pp. 2153–2157.
 12. M. Dredze, "How Social Media Will Change Public Health," *IEEE Intelligent Systems*, vol. 27, no. 4, 2012, pp. 81–84.
 13. M. Salathé et al., "Digital Epidemiology," *PLoS Computational Biology*, vol. 8, no. 7, 2012; doi:10.1371/journal.pcbi.1002616.
 14. J. Ginsberg et al., "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature*, vol. 457, Feb. 2009, pp. 1012–1014.
 15. V. Lampos and N. Cristianini, "Nowcasting Events from the Social Web with Statistical Learning," *ACM Trans. Intelligent Systems and Technology*, vol. 3, no. 4, 2012; doi:10.1145/2337542.2337557.
 16. A. Lamb, M. Paul, and M. Dredze, "Separating Fact from Fear: Tracking Flu Infections on Twitter," *Proc. North Am. Chapter Assoc. Computational Linguistics Conf.*, Assoc. Computational Linguistics, 2013, pp. 789–794.
 17. M. Paul and M. Dredze, "You Are What You Tweet: Analyzing Twitter for Public Health," *Proc. 5th AAAI Int'l Conf. Weblogs and Social Media*, AAAI, 2011, pp. 265–272.
 18. M. Salathé and S. Khandelwal, "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control," *PLoS Computational Biology*, vol. 7, no. 10, 2011; doi:10.1371/journal.pcbi.1002199.
 19. A. Sadilek, H. Kautz, and V. Silenzio, "Modeling Spread of Disease from Social Interactions," *Proc. 6th AAAI Int'l Conf. Weblogs and Social Media*, AAAI, 2012; www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4493/4999.
 20. N. Christakis and J.H. Fowler, "Social Network Sensors for Early Detection of Contagious Outbreaks," *PLoS ONE*, vol. 5, no. 9, 2010; doi:10.1371/journal.pone.0012948.
 21. H. Shao et al., *Predicting the Flu before It Happens: Designing Social Network Sensors for Epidemics*, tech. report 13-063, Network Dynamics and Simulation Science Laboratory (NDSSL), Virginia Bioinformatics Inst., Virginia Tech, 2013.
 22. T. Lengauer and R. Tarjan, "A Fast Algorithm for Finding Dominators in a Flowgraph," *ACM Trans. Programming Languages and Systems*, vol. 1, no. 1, 1979, pp. 121–141.
 23. A. Apolloni et al., *Optimal Vaccine Allocation and Vulnerability*, tech. report 10-504, Network Dynamics and Simulation Science Laboratory (NDSSL), Virginia Bioinformatics Inst., Virginia Polytechnic Inst. and State Univ., 2010.
 24. M. Becker, ed., "The Health Belief Model and Personal Health Behavior," *Health Education Monographs*, vol. 2, no. 4, 1974, pp. 324–508.
 25. A. Bandura, *Social Foundations of Thought and Action: A Social Cognitive Theory*, Prentice Hall, 1986.
 26. S. Funk, M. Salathé, and V. Jansen, "Modelling the Influence of Human Behaviour on the Spread of Infectious Diseases: A Review," *J. Royal Soc. Interface*, vol. 7, no. 50, 2010, pp. 1247–1256.
 27. K. Bisset et al., "Indemics: An Interactive Data Intensive Framework for High Performance Epidemic Simulation," *Proc. 24th ACM Int'l Conf. Supercomputing*, ACM, 2010, pp. 233–242.
 28. J. Chen, A. Marathe, and M. Marathe, "Coevolution of Epidemics, Social Networks, and Individual Behavior: A Case Study," *Advances in Social Computing*, LNCS 6007, Springer, 2010, pp. 218–227.

Madhav V. Marathe is a professor of computer science and the deputy director of the Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech. His research interests include network science, computational epidemiology, high-performance computing, and policy informatics. Marathe has a PhD in computer science from University at Albany, State University of New York. He is an IEEE Fellow. Contact him at mmarathe@vbi.vt.edu.

Naren Ramakrishnan is the Thomas L. Phillips Professor of Engineering at Virginia Tech and director of the university's Discovery Analytics Center. His research interests include knowledge discovery from datasets arising in health informatics, sustainability, and intelligence analysis. Ramakrishnan has a PhD in computer sciences from Purdue University. Contact him at naren@cs.vt.edu.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.