An Interaction-Based Approach to Computational Epidemiology

Christopher L. Barrett, Stephen Eubank and Madhav V. Marathe *

Introduction

Epidemiology is the study of patterns of health in a population and the factors that contribute to these patterns. Computational Epidemiology is the development and use of computer models to understand the spatio-temporal diffusion of disease through populations. An important factor that greatly influences an outbreak of an infectious disease is the structure of the interaction network across which it spreads. Aggregate or collective computational epidemiology models that have been studied in the literature for over a century, often assume that a population is partitioned into a few subpopulations (e.g. by age) with a regular interaction structure within and between subpopulations. Although useful for obtaining analytical expressions for a number of interesting parameters such as the numbers of sick, infected and recovered individuals in a population, it does not capture the complexity of human interactions that serves as a mechanism for disease transmission. In other words, the aggregate approach does not take the structure of underlying social network into account. Additionally, the number of different subpopulation types considered is small and parameters such as mixing rate and reproductive number are either unknown or hard to observe.

Here we describe *Simdemics*: an interaction-based multiagent approach to support epidemic planning for large urban regions. *Simdemics* is an example of a disaggregated modeling approach in which interactions between every pair of individuals is represented. It is based on the idea that a better understanding of the characteristics of the social contact network can give better insights into disease dynamics and intervention strategies for epidemic planning.

Simdemics details the demographic and geographic distributions of disease and provides decision makers with information about (1) the consequences of a biological attack or natural outbreak, (2) the resulting demand for health services, and (3) the feasibility and effectiveness of response options. A unique feature of *Simdemics* is the size and scale of urban regions that can be analyzed using it.

*Network Dynamics and Simulation Science Laboratory, Virginia Bio-Informatics Institute, Dept. of Computer Science, Virginia Tech, {cbarrett,seubank,mmarathe}@vbi.vt.edu

Simdemics uses a number of concepts studied in traditional and contemporary AI literature. This includes, multiagent systems, social network analysis, Markov decision processes and large n-way games [4, 8, 11]. However, the practical use of this tool prompted the investigation of several new basic and applied research questions. For example, we had to develop new HPC oriented efficient algorithmic techniques to *generate and analyze* dynamic social networks and *simulate diffusion of diseases* on these dynamic social networks [13, 20]. These algorithms were implemented so that they can scale to 100 million node networks and can be mapped on to 100-1000 processor shared memory multiprocessor architectures [8, 13, 20]. Similarly, scalable data mining methods are being developed to analyze the vast data sets that are produced by *Simdemics* [11]. These scalable simulations and mining algorithms form the basis of practical and usable decision support systems that we have built and are being continually enhanced [1, 3].

Basic Approach

The overall approach consists of composing four distinct models: **Step 1.** Model for creating a set of synthetic individuals, **Step 2.** Model for generating a (time varying) interaction networks, **Step 3.** Model for simulating the epidemic process, and **Step 4.** Model for representing and evaluating interventions and public policies. The overall mathematical model consists of two parts: (i) a discrete dynamical system framework that captures the co-evolution of disease dynamics, social network and individual behavior (first three steps) and (ii) a partially observable Markov decision process that captures various control and optimization problems formulated on the phase space of this dynamical system. See [13, 20] for more details.

Step 1 creates a synthetic urban population by integrating a variety of databases from commercial and public sources into a common architecture for data exchange. The process preserves the confidentiality of the original data sets, yet produces realistic attributes and demographics for the synthetic individuals. The synthetic population is a set of synthetic people and households, located geographically, each associated with demographic variables drawn from any of the demographics available in the census. Joint demographic distributions can be reconstructed from the marginal distributions available in typical census data using an iterative

proportional fitting (IPF) technique. Each synthetic individual is placed in a household with other synthetic people and each household is located geographically in such a way that a census of our synthetic population yields results that are statistically indistinguishable from the original census data, if they are both aggregated to the block group level.

In Step 2, a set of activity templates for households are determined, 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. Thus for a city - demographic information for each person and location, and a minute-by-minute schedule of each person's activities and the locations where these activities take place is generated by a combination of simulation and data fusion techniques. This yields 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 points. It is *impossible* to build such a network by simply collecting field data; the use of generative models to build such networks is a unique feature of this work.

Step 3 consists of developing a computational model for representing the disease within individuals and its transmission between them. The model can be viewed as a *networked probabilistic timed finite state machine*. Each individual is associated with a timed probabilistic finite state machine. Furthermore, the automata are connected to other automata – this coupling is derived from the social contact network. The state transition is probabilistic and is timed (i.e. depends on the duration of contact). It may also depend on the attributes of the people involved (age, profession, health status, etc.) as well as the type of contact (intimate, casual, etc.). states of individual automata as they update their states in responses to changes in internal state and state of its neighbors.

Step 4 consists of representing and analyzing various public policies and interventions using a combination of partially observable Markov decision process (POMDP) and n-way games. It allows us to capture sequential decision making process related to studying the efficacy of various interventions and behaviors of individual agents in response to their perception of disease spread. The POMDP is exponentially larger than the problem specification and is intractable to solve optimally in general. We thus resort to efficient simulations. A key concept is that of *implementable policies* — policies or interventions that are implementable in the real world.

From Theory to Practice and Back to Theory

Simdemics is being developed continually over the last 12 years. It was used in a number of user defined studies, including recent pandemic planning studies undertaken for DHS, DoD and DHHS. The studies have guided the continued evolution of *Simdemics*. Equally important, these studies helped us identify new research questions at the interface of multi-agent modeling, data mining, network science and high performance computing. For e.g., recently, at the request of federal agencies involved in preparing for an influenza pandemic, *Simdemics* as a part of NIH funded MIDAS group analyzed combinations of strategies for responding to influenza. Results of the MIDAS analysis were reviewed in a Letter Report by the Institute of Medicine, Modeling Community Containment for Pandemic Influenza [23]. We discuss this below.

The MIDAS study considered both pharmaceutical and non-pharmaceutical interventions (NPI) targeted at those parts of the population where they might most effectively control the spread of disease. NPIs aim to alter human social behaviors so as to mitigate an outbreak. They include interventions such as closing schools or reducing contacts at work and in the community [1, 23]. In the course of this study, we found our overall methodology to be suitable for estimating normal social contact patterns as well as changes in patterns resulting from NPIs [23]. Indeed, it is difficult to generalize observations about transmission in observed outbreaks to hypothetical circumstances without such a generative, structurally calibrated model of social networks. For example, one can estimate from historical outbreaks that roughly 35% of influenza transmission occurs within a household and 65% occurs in the community (e.g. at work, school, etc.). However, if the proportion of transmission occurring in different contexts is a parameter of the model, it becomes impossible to say how it might change as people's behaviors change. Moreover, since it is difficult to find out what spontaneous changes in behavior were happening during the historical outbreaks, it may be that the effects of NPIs on overall transmission patterns are already included in transmission parameter estimates. Our methods infer the proportions given a social network from assumptions about relative transmission rates between people with different demographics, in effect separating the problem of estimating the social network from the problem of estimating transmission over that network. The MIDAS and other recent studies have raised a number of new (and sometimes new twists on old) research questions of interest to the AI community. We give two examples of this below. The examples are discussed in the context of public health epidemiology, but can often be generalized to other socio-technical systems.

(a) Dynamic Graphical Models: Multi-agent models to study co-evolution of large social networks, individual behavior, disease dynamics and public policies. Over the last several years, researchers in AI have been studying the interaction between individual actions and the multi-agent networks that they constitute. Graphical models of Bayesian inference and games have been proposed and studied in AI to capture the network structure inherent in certain applications. Epidemics on social networks provides a useful and realistic application for further studying graphical games and inference problems [17, 21]. The inclusion of public policies and disease dynamics as a part of this interaction process motivates several new questions, ranging from representation of realistic agent behaviors in crises [1, 3, 10, 19] to design of heuristic methods for solving op-

timization problems modeled as a combination of POMDP and n-way games to data mining methods for inferring spatial and temporal patterns of disease spread to assist interventions [11, 18]. A single case study involved addressing variants of all of the above questions. The size of the systems (10-300 million agents) makes the problems even more challenging. The recent work on computational methods for analysis of graphical models has shown how tree-like structures can be exploited to compute well known quantities, e.g. Nash equilibria [17, 22, 26]. Unfortunately, social networks of urban regions do not have the tree-like property. They are not even small-world networks or scale free networks as defined in the current literature; see [8, 12]. Understanding the structure of these networks and exploiting this structure for designing efficient computational solutions is an important research question. Another interesting direction for further research is to extend the notion of graphical games and inference problems to *dynamic graphical models* — games and inference problems in which the underlying network is changing due to the decisions taken by individual agents.

(b) Intelligent query processing systems and computational steering of simulation-based experiments. This topic is related to classical problems in AI and is being revisited in the context of semantic web and knowledge-based systems [6]. Other efforts, such as the The UK E-Science initiative is concerned with many of the same questions. As we started using our models to address user defined questions, it became progressively clear to us that easy to use web-based systems would provide an appropriate mechanism for delivering the results obtained by executing our models. Our goal is to allow the analyst who is not a computing expert, to use HPC-based models routinely and with ease. Given that the underlying complex network, individual behavior and dynamics of particular process over the network (e.g. epidemic) coevolve, we require an adaptive computational steering mechanism. This requires methods for coordinating resource discovery of computing and data assets; AI-based techniques for translating user level request to efficient workflows; reusing data sets whenever possible and spawning computer models with required initial parameters and coordination of resources among various users. Consider a hypothetical yet illustrative query by an analyst: *Compare the effects of vaccinating 10% of the school children or closing schools two days to control a potential flu epidemic city of Portland and its surrounding areas.* In order to answer this simple query, the system will spawn a series of sub-queries and computational tasks, e.g. does such a data already exist in our database ? It might be that the data exists but with 12% of school children being vaccinated in the study. The system then needs to determine if the results are *adequate*. If the answer is no, then we might potentially have to create the Portland social network, construct an appropriate experimental design that consists of choosing various subsets of school children (note that the query does not specify this precisely and hence one might consider a random or the optimal subset), and so on. Statistical analysis of the results will then be performed and the final results presented to the analyst as a combination of charts, spatial spread movies using Google Earth and so on. We have only taken the first steps in addressing both these research areas, see [2] for additional discussion.

Acknowledgments. We thank our external collaborators and members of the Network Dynamics and Simulation Science Laboratory(NDSSL); the work reported here is a joint effort of all the team members. This work has been partially supported NSF Nets Grant CNS-062694, HSD Grant SES-0729441, CDC Center of Excellence in Public Health Informatics Grant 2506055-01, NIH-NIGMS MIDAS project GM070694-06, DTRA CNIMS Grant HDTRA1-07-C-0113.

References

[1] K. Atkins et al. Simulated Pandemic Influenza Outbreaks in Chicago VT *TR-NDSSL-07-004*, 2004.

[2] K Atkins, C. Barrett, R. Beckman, K. Bisset, J. Chen, S. Eubank, A. Feng, Z. Feng, S. Harris, B. Lewis, V. Anil Kumar, M. Marathe, A. Marathe, H. Mortveit, P. Stretz, An Interaction Based Composable Architecture for Building Scalable Models of Large Social, Biological, Information and Technical Systems, *CTWatch Quarterly*, Volume 4, Number 1, March 2008.

[3] K. Atkins, et al. An analysis of layered public health interventions at Ft. Lewis and Ft. Hood during a pandemic influenza event VT *TR-NDSSL-07-019*, 2007.

[4] C. Barrett, S. Eubank, V. Anil Kumar, M. Marathe, Understanding Large Scale Social and Infrastructure Networks: A Simulation Based Approach, *SIAM news: The Mathematics of Networks*,2004.

[5] C. L. Barrett, K. Bisset, S. Eubank, V. S. A. Kumar, M. V. Marathe and H. S. Mortveit, Modeling and Simulation of Large Biological, Information and Socio-Technical Systems: An Interaction-Based Approach, *Proc. Symposia in Applied Mathematics, Short Course on Modeling and Simulation of Biological Networks, AMS Lecture Notes, Series, (PSAPM),* 64, pp. 101-147, 2007.

[6] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, *Scientific American*, May (2001).

[7] C. Barrett, K. Bisset, J. Chen, B. Lewis, S Eubank, V.S. Anil Kumar, M. Marathe, H. Mortveit, Effect of Public Policies and Individual Behavior on the Co-evolution of Social Networks and Infectious Disease Dynamics, *Proc. DIMACS/DyDAn Workshop on Computational Methods for Dynamic Interaction Networks*, September 2007.

[8] C. Barrett, S. Eubank and M. Marathe Modeling & Simulation of Large Biological, information and Socio-Technical Systems: An Interaction Based Approach, *Interactive Computing: A new Paradigm*, Ed. D. Goldin, S. Smolka and P. Wegner pp. 353-394, Springer Verlag, 2006.

[9] C. Barrett, S. Eubank, J. Smith, If smallpox strikes Portland ... *Scientific American,* 292, 2005.

[10] C.T. Bauch and D.J. Earn, Vaccination and the theory of games, *Proc. Natl. Acad. Sci.*, 101(36), pp. 13391-13394, 2004.

[11] C. Bailey-Kellog, N. Ramakrishnan, M. Marathe, Spatial data mining to support pandemic preparedness. *SIGKDD Explorations* 8: 80-82, 2006.

[12] S. Eubank, V.S. Anil Kumar, M. Marathe, A. Srinivasan and N. Wang, Structure of Social Contact Networks and Their Impact on Epidemics, *AMS-DIMACS Special Issue on Epidemiology,* 70, pp. 181-213, 2006.

[13] S. Eubank, et al. H. Guclu, V.S. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai and N. Wang. Modeling Disease Outbreaks in Realistic Urban Social Networks. *Nature,* 429, (2004).

[14] N.Ferguson, D. A. T. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, D. S. Burke Strategies for mitigating an influenza pandemic, *Nature*, April, 2006.

[15] N. L. Ferguson, D. A. T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Lamsirithaworn, D. S. Burke Strategies for containing an emerging influenza pandemic in Southeast Asia, *Nature*, vol 437, September, 2005.

[16] T.C. Germann, K. Kadau, I. M. Longini Jr., C. A. Macken Mitigation strategies for pandemic influenza in the United States, *Proc. of National Academy of Sciences (PNAS)*, April 11, vol 103, no. 15, pp. 5935-5940, 2006.

[17] M. Kearns *Graphical Games, Algorithmic Game Theory*, Ed. N. Nisan, T. Roughgarden, E. Tardos and V. Vazirani, Cambridge University Press, pp. 159-180, 2007

[18] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the Spread of Influence in a Social Network, *Proc. KDD,* (2003).

[19] J. Epstein, J. Parker, D. Cummings. Coupled Contagion Dynamics of Fear and Disease: A Behavioral Basis for the 1918 Epidemic Waves: Mathematical and Computational Explorations Technical report, Brookings Institute. Presentation made at the MIDAS meeting, June 2006.

[20] S. Eubank, et al. V.S. Anil Kumar, M. Marathe, A. Srinivasan and N. Wang. Structural and Algorithmic Aspects of Large Social Networks, *Proc. 15th ACM-SIAM Symposium on Discrete Algorithms (SODA),* pp. 711-720, (2004).

[21] S. Lauritzen, *Graphical Models* Oxford University Press, 1996.

[22] S. Lauritzen and D. Spiegelhalter, Local Computation with probabilities on graphical structures and their application to expert systems, *J. Royal Statistical Society B* 50(2), pp. 157-224, 1988.

[23] Letter Report (2007) National Academies Press `http://www.nap.edu/catalog/11800.html`

[24] M. Mundhenk, J. Goldsmith, C. Lusena and E. Allender, Complexity of finite-horizon Markov decision process problems. *J. ACM (JACM)* 47(4), 2000, pp. 681-720.

[25] M. E. J. Newman. *The structure and function of complex networks*. SIAM Review **45**, 167–256 (2003).

[26] J. Pearl, *Probabilistic Reasoning in Intelligent Systems,* Morgan Kaufmann, 1988.

[27] D. Vickery and D. Koller Multi-agent algorithms for solving graphical games, *Proc. 18th International Conference on Artificial Intelligence (AAAI)*, pp. 345-351, 2002.

Brief Biography

Madhav Marathe is a Professor, of Computer Science, and Deputy Director of Network Dynamics and Simulation Science Laboratory (NDSSL), Virginia Bio-Informatics Institute at Virginia Polytechnic Institute and State University. He obtained his B.Tech degree in 1989 in Computer Science and Engineering from the Indian Institute of Technology (IIT) Madras, and his Ph.D. in 1994 in Computer Science, from University at Albany under the supervision of Professors Harry B. Hunt III and Richard E. Stearns. Before coming to Virginia Tech, he was a Team Leader in the Basic and Applied Simulation Science group (CCS-5) in the Computer and Computational Sciences division at the Los Alamos National Laboratory (LANL) where he led the theoretical program to support simulation based design, and analyze extremely large socio-technical and critical infrastructure systems. At Los Alamos, he played a lead role in the routing module of Transportation Simulation and Analysis System (TRANSIMS), AdHopNet – a modeling tool for integrated advanced communication networks, and a team leader for the Urban Infrastructure Suite (UIS), funded as part of the the DHS National Infrastructure Simulation and Analysis Center (NISAC). Since joining Virginia Tech, he and NDSSL team members have been actively developing an interaction based composable architecture for building scalable models of large socio-technical systems; see [2]. He serves as a Co-PI/PI/Senior Investigator on a number of projects related to computational epidemiology. This includes: NIH-MIDAS project for developing agent based model to represent and analyze spread of infectious diseases (Co-PI), CDC Center of Excellence in Public Health Informatics (University of Utah Medical School is the lead) aiming to develop innovative methods and informatics infrastructure to support epidemic preparedness, (Co-PI), NSF HSD program to study the effect of individual behaviors on the evolution on epidemics, (PI).