Ubiquity Symposium

Big Data

High Performance Synthetic Information Environments

An integrating architecture in the age of pervasive data and computing

by Christopher L. Barrett, Jeffrey Johnson, and Madhav Marathe

Editor's Introduction

The complexities of social and technological policy domains, such as the economy, the environment, and public health, present challenges that require a new approach to modeling and decision-making. The information required for effective policy and decision making in these complex domains is massive in scale, fine-grained in resolution, and distributed over many data sources. Thus, one of the key challenges in building systems to support policy informatics is information integration. Synthetic information environments (SIEs) present a methodological and technological solution that goes beyond the traditional approaches of systems theory, agent-based simulation, and model federation. An SIE is a multi-theory, multi-actor, multi-perspective system that supports continual data uptake, state assessment, decision analysis, and action assignment based on large-scale high-performance computing infrastructures. An SIE allows rapid course-of-action analysis to bound variances in outcomes of policy interventions, which in turn allows the short time-scale planning required in response to emergencies such as epidemic outbreaks.

Ubiquity Symposium

Big Data

High Performance Synthetic Information Environments

An integrating architecture in the age of pervasive data and computing

by Christopher L. Barrett, Jeffrey Johnson, and Madhav Marathe

Many public policy issues nowadays involve biology, information, sociology, and technology (BIST) systems, which consist of large numbers of interacting physical, biological, technological, informational and human/societal components whose global system properties are a result of interactions among representations of local system elements. In other words, the behavior of each component and the interactions among components affect the outcome at the system, as well as global, level. At the same time, the behavior and interactions of elements are affected by the global state. The interdependencies between individual components and the dynamical interactions affect the global outcome. The feedback between dynamics and the structure is bidirectional and makes these systems difficult to control. Also, they involve multiple stakeholders, who often have conflicting optimization criteria. These interactions and interdependencies can be abstracted representationally as networks, and network and data science provide a framework to explicitly and intuitively model local interactions and analyze the outcomes from a global perspective. There is a rich literature on studying networked systems of interest to public policy, including urban regional transportation systems, national electrical power markets and grids, ad hoc communication and computing systems, and public health systems.

The co-evolution of network structures and the local interactions in each network often result from individual decision-making processes, and understanding them requires a detailed and systematic modeling approach. Traditional modeling methods fall short, considering the complexity of the problems involved. Simplifying assumptions, made to ensure tractability of analysis, often reduce the validity and applicability of models.

A novel solution for present-day policy informatics problems has to provide (a) support for multiple views and multiple optimization criteria, for the multiple stake-holders (adaptability); (b) the capability to incorporate multiple sources of data (extensibility); (c) the capability to

model very large, interacting, networked systems (scalability); and (d) support for policy planning by enabling the evaluation of a large class of possible interventions (flexibility). The team at the Network Dynamics and Simulation Science Laboratory (NDSSL) at Virginia Tech's Biocomplexity Institute (BI) has developed a methodology to integrate information from multiple sources and to build large-scale high-resolution simulations to address these challenges. Next, we describe our approach and how it satisfies the four properties mentioned above.

Synthetic Information Systems for Information Integration and Management

Digital data are being generated at an amazing rate. All the digital data generated through computers, mobile phones, digital cameras, television, etc., has us entering the yottabyte (10^{24}) era. Various large-scale surveys provide additional data such as census and consumer behavior. It is tempting to think that the informatics problem is simply to organize all the data available and extract the information we need to solve our problems, especially when we have such unprecedented data sources. While extracting information from such massive amounts of data would be challenging in itself, often the real problem is we do not have the right data for the problem at hand. Available observations and knowledge are normally not structured specifically for a particular question. We overcome this data problem by using available, sometimes imperfect, information in the form of data and procedures, to synthesize an integrated representation of what is known in the context of the decision to be made.

From an informatics perspective, there are two important things to note about this approach. First, it goes beyond traditional informatics notions of indexing and mining, by combining many sources of data into a model that encodes nominative, declarative, and procedural knowledge. In addition to policy planning and simulations, this allows consistency checks of the data sources, and also exposes gaps in data, which can guide future data collection efforts. Therefore, we call this approach "model-based informatics."

Synthetic Agent, Populations, and Networks

A synthetic agent is a representation of the elements and states of an agent to provide a statistically accurate overall picture, not to precisely match any snapshot of an agent. An agent here can refer to people, places, things, cells, cytokines, organs, autonomous agents, etc. For example, a synthetic human agent can have demographic, social, health, cognitive, cultural

attributes. These need to be statistically accurate to human attributes. The associated data are derived from real-world measurements including:

- *nominative data*, e.g. age, income, gender, physiological data such as height, blood group, genome, etc.
- *declarative data* such as specific activities persons might perform, and described as a list or vector of objects
- *procedural data* that includes how they might respond to external events and is best described as an algorithm or a method

The term “synthetic” is used in two ways: (i) data and attributes of the human are synthesized by integrating a diverse set of data sources and using models for interpolation and extrapolation of data, and (ii) a synthetic agent is “similar” to real individuals but is not identical to any individual in the population.

A synthetic agent comprises of individual attributes based on real-world collected data. Privacy of individuals is protected. The correlations between the data sets agree with the measured correlations of data in the real world, e.g. if I say there is a synthetic individual whose height is 20 feet then this is unlikely based on the existing data. In other words, a synthetic human is statistically similar to a group of individuals in the society, but is not identical to any of them.

Our work in inferring large-scale functioning societal infrastructures and populations is summarized in [1, 2, 3]. When taken in aggregate the synthetic persons (agents) and their activities form a social contact network where individuals come in contact with other individuals in the synthetic population at specific times and locations. Understanding the properties of these social contact networks is essential in understanding what people are doing, where they are doing it, and the consequences of policy decisions, such as evacuations in the cases of man-made or natural disasters or sub-population vaccinations in the case of pandemic influenza. For example, examination (confirmed by simulation) of a social contact network of an urban area shows an effective distribution scheme of a limited supply of vaccines is to vaccinate school children. The detail and theory used in the construction of the above social contact networks is essential. In contrast to several recent results, we have shown activity-based social contact networks generated as above differ from classical models of random networks. In particular, we show social contact networks for U.S. cities—such as Portland, Chicago, and Houston—have the following distinguishing features: (i) they are neither scale-free nor small-world networks (contrary to widespread belief); and (ii) they have high local clustering, while

most physical networks have low clustering coefficients. Hence, decisions based on relative simplistic networks, such as random networks, have a large probability of being incorrect.

In the above description of the construction of the social contact networks little was said on the “why” some synthetic individual is doing what it is doing. In some special cases, where behavioral changes are obvious, we have added some dynamic behavioral components to the social contact structures. That is, an individual assesses the state of the system and changes its behavior and hence its activity sequences and the corresponding social contact network. Examples of this include staying home from work or school when sick during a pandemic influenza, closing schools, and changed activity sequences of members of the National Guard when called to service.

Dynamic agent-based models are then designed to study the dynamical evolution phenomenon of interest. Our agents are numerous (e.g. while studying pandemics we need to study the entire U.S. population comprising of 300 million agents) combined with detailed representation of the underlying infrastructure networks. The scale and scope of the models imply high performance computing (HPC) oriented methods are essential to achieve the needed computational efficiency. The research proposes developing agent-based models to study behavioral modeling questions such as: “What makes societies collapse into disorder, including violence?”, “What makes previously unstable, transitional societies cohere?”, “How does the contagion of fear spread through a society?”, “How do ideas and beliefs flow through a population?”, and “How might data on media use be used to understand the social structure of influence in a population?” For example, it is likely that in some populations certain patterns of phone use would be indicative of a close (influential) relationship, whereas in others the only powerful indicator would be actual face-to-face contact. We evaluate the extent to which behavioral data, such as e-mail, instant messaging, collocation, or mobile and landline phone calls might be used as indicators of social influence.

A multi-theory approach is required to study the problem. Generalized threshold models can be used to model contagion (influence), but, in addition, we need co-evolutionary models to represent how the network evolves and persists as a result of the contagion. The network and the contagion are coupled and co-evolve and modeling this dynamical process is one of the important research questions that we are currently studying. Our models do not make classical assumptions that agents are rational and forward-looking, respond promptly to changes in their environment, or have and use extensive information about their situation. Furthermore, we assume agents are heterogeneous—there are differences in demographics, tastes, location, experience, and rationality in how each agent perceives and interacts with other agents and the environment. Traditionally heterogeneity is acknowledged but generally not modeled in social

sciences. Our agents adapt to local conditions based on fragmentary information, the behavior of others in their social networks, and external shocks.

The Role of High Performance Computing (HPC)

HPC will play a crucial role in developing real-time decision support, information acquisition and analytical environments for GSS.

Scaling and effective utilization of supercomputing resources. Scaling to large machines and large instances so as to complete computations in a reasonable time is necessary and now feasible. Supercomputing resources will be critical for modeling global scale systems at detailed spatial, temporal and individual levels. A simple calculation suggests an individual-based representation of such global scale networks will to have 10^{9-11} agents with 10^{11-14} edges. Structural analysis of these networks as well as dynamics over such networks motivates the use of current and emerging supercomputing resources. Developing models that can effectively use supercomputing platforms is challenging. The networks are highly irregular, dynamic and co-evolving. The emerging petascale and future exascale computing platforms will have 1 million - 100 million+ cores. As an example of recent progress, we submitted a paper to CCGrid Conference that shows for the first time how social simulations (we did this for epidemics in the paper) can be mapped onto machines with more than 300,000 cores. This is the largest open machine in the U.S. at the National Center for Supercomputing Applications (NCSA). We can now run a single run of epidemic simulation (200 days) for the entire U.S. in about 10 seconds. The network has 300 million nodes and 15 billion+ edges.

Global Systems Science (GSS) for policy will be a major user of the technologies discussed. GSS has been characterized as a combination of policy problems, complex systems science, policy informatics, and citizen engagement [4]. It anticipates the use of massive social simulations to address interdisciplinary policy questions at local and global scales. Scaling such as described here will become critical as we move to developing detailed models for GSS. The scaling we are getting to process structural properties (not dynamics) is even better. Our goal is to have models that scale to 10 billion node networks in about three years. This will get us ready for addressing the important questions raised in the GSS program and supporting real-time policy making.

HPC-enabled methods for data analytics. The kinds of models we would like to develop for GSS should be driven by a combination of data and appropriate theories. Data here is meant to refer to classical use of data, but also procedural information in the form of laws, behaviors and

policies, as well as networked relationships that capture interactions, causality, and dependencies. HPC methods are therefore needed to process these data sets to prepare them as input for dynamic models. The resulting data, all of which are really a part of the synthetic information, should also be processed to identify important patterns, trends, anomalies, etc. HPC hardware and methods for this are often quite different than the traditional clusters used to run large models. A recent trend is the concept of data intensive supercomputing. The new model of computation differs from traditional models of computing in that producing, analyzing, processing, and curating data are integral parts of the computation. Big data is a related concept and focuses on related concepts, including analytics and reasoning. We have recently proposed the concept of “network centric computing” that extends these ideas. It melds the traditional data- and compute-intensive approaches and also highlights the role of networked data. In network-intensive computing, HPC resources are used to compute about and over networks. Moreover, the computation requires significant amount of data to synthesize the networks as well as significant computing to process these networks.

Massively distributed data collection & computing. Crowdsourcing, pervasive availability of devices, and sensor systems all point to the need for a different notion of HPC. In this view we are talking about highly distributed, fault tolerant, spatially distributed, “bursty” data and computation. To support citizen politics and decision-making, as well as real-time data gathering, this form of computing will become increasingly important in the coming years. Crowdsourcing of computation can occur at various levels— from simple collection and dissemination of information and data to active computation in which humans are a part of a distributed computing process. Crowdsourcing has played an important role in policy making and citizen science already. For example, the role of social media and crowd-sourced methods was evident most notably during the recent social revolutions in the Middle East under the rubric of the Arab Spring.

Distributed real-time decision-making. The availability of data at very fine scales (temporal, spatial, social) is prompting individuals, groups and organizations to develop real-time, decision-making abilities. This includes, rapid changes in how resources are brought to bear on a problem, as well as interventions that are analyzed and enacted to reduce the severity of the problem. Of course the time scales of policy making in the past are quite different. One can already see glimpses of this. Responses to market crashes, pandemics, and natural disasters create a tension between the expectations of the public at large and businesses, governments, and institutions. Decision-making in such settings is always done with incomplete information and the system is co-evolving with the decisions. Thus appropriateness of decisions will be questioned: “Was the response to H1N1 pandemic too slow or too fast?”, “Was the response

too aggressive?”, “Was the response to the volcanic eruption over Finland too slow?”, etc. GSS will need to address how to make faster decisions, how to analyze the massive amounts of data and study the possible counter-factuals, and how to convey these decisions to the public.

Empowering citizens to be decision makers. An important outcome of today’s globally connected world is that individuals, small communities, and organizations can participate in the entire decision-making process in a manner that was not possible before. This changes the dynamics of global systems, which were traditionally managed by centralized and hierarchical authorities. GSS will need to address and develop protocols and information sharing schemes for networked decision-making. This includes methods for allowing individuals to convey their preferences, thoughts, votes, and ideas to traditional decision makers. It also includes the need for methods to make information related to the event available to individual decision makers; and creating online tools for them to effectively interact with other individuals.

Conclusions

Model-based informatics and synthetic information environments (SIEs) have been presented as an approach to synthesizing massive heterogeneous, but usually, inappropriately structured data sources into an integrated form to provide the information required for effective policy and decision making for complex domains at local and global scales. Synthetic agents provide a statistically accurate overall picture of all the agents in a system, while not precisely matching any of them. Thus the properties of all the agents are captured without compromising anonymity and privacy. The data are derived from real-world measurements, including: nominative data such as age, income and gender; declarative data such as the specific actions a person might perform; and procedural data such how agents may respond to external events. Dynamic models are used to study the phenomena of interest, such as the behavior of people during a pandemic flu, or the possible impacts of policy, such as building new transportation infrastructure. HPC plays an essential role in model-based informatics, and is essential for major uses such as GSS for policy at all scales. Crowdsourcing, pervasive availability of devices and sensor systems all point to the need for a different approach to HPC. An important outcome of these developments is empowering citizens to participate in decision-making in ways that was not possible before.

Acknowledgements

The work described here is done jointly with member of the Network Dynamics and Simulation Science Laboratory. The authors were supported partially by NSF EAGER grant, DARPA NGS2, DTRA CNIMS, NSF BIG DATA, NSF DIBSS grants.

References

- [1] C. Barrett, S. Eubank, and M. Marathe. Modeling and Simulation of Large Biological, Information and Socio-Technical Systems: An Interaction Based Approach. In *Interactive Computation: The New Paradigm*, Springer Verlag, 2005.
- [2] C. Barrett, S. Eubank, V. Anil Kumar, and M. Marathe. Understanding Large Scale Social and Infrastructure Networks: A Simulation Based Approach. *SIAM News* (March 2004). Appears as part of Math Awareness Month on The Mathematics of Networks.
- [3] S. Eubank, H. Guclu, V.S. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modeling Disease Outbreaks in Realistic Urban Social Networks. *Nature* (2004).
- [4] Dum R., and J. Johnson. Global Systems Science and Policy. In *Non-Equilibrium Social Science and Policy. Understanding Complex Systems*. Springer, 2017.

Further Reading

- Capturing Complexity through Agent-Based Modeling. *PNAS* Special Issue, 2002.
- R. Albert and A. Barabasi. Statistical mechanics of complex networks, *Rev. Mod. Phys.* 74 (2002), 47-97.
- A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 1999.
- 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 In *Proc. Short Course on Modeling and Simulation of Biological Networks*. AMS Lecture Notes. Series: PSAPM. 2007.
- C. Barrett, S. Eubank and M. Marathe. Modeling and Simulation of Large Biological, Information and Socio-Technical Systems: An Interaction Based Approach. In *Interactive Computation: The New Paradigm*, D. Goldin, S. Smolka and P. Wegner Eds. Springer Verlag, 2005.
- C. Barrett, J.P. Smith and S. Eubank. Modern Epidemiology Modeling. *Scientific American* (March 2005).

C. Barrett, R. Beckman, K. Berkbigger, K. Bisset, B. Bush, K. Campbell, S. Eubank, K. Henson, J. Hurford, D. Kubicek, M. Marathe, P. Romero, J. Smith, L. Smith, P. Speckman, P. Stretz, G. Thayer, E. Eeckhout, and M.D. Williams. TRANSIMS: Transportation Analysis Simulation System. Technical Report LA-UR-00-1725. Los Alamos National Laboratory. Unclassified Report, 2001. An earlier version appears as a seven-part technical report series LA-UR- 99-1658 and LA-UR-99-2574 to LA-UR-99-2580.

R. Breiger and K. Carley, Eds. NRC Workshop on Social Network Modeling and Analysis. National Research Council, 2003, 133-145.

C. Barrett, S. Eubank and M. Marathe. Modeling and Simulation of Large Biological, Information and Socio-Technical Systems: An Interaction Based Approach. In *Interactive Computation: The New Paradigm*, D. Goldin, S. Smolka and P. Wegner (Eds.). Springer, 2005.

J. Carlson, and J. Doyle. Complexity and robustness. *Proc. National Academy Science (PNAS)* 99, (2002), 1317-1345.

Grid Computing: Making the Global Infrastructure a Reality, Fran Berman, Geoffrey Fox and Tony Hey (Eds.). Wiley Publishers, March 2003.

C. Barrett et al. Reachability Problems for Sequential Dynamical Systems with Threshold Functions. *Theoretical Computer Science* 1-3, (2003), 41-64.

C. Barrett et al. Complexity of Reachability Problems for Finite Discrete Sequential Dynamical Systems. *J. Computer and System Sciences* 72, (2006), 1317-1345.

V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. Prediction and predictability of global epidemics: the role of the airline transportation network. In *Proceedings of the National Academy of Sciences*. 2006, 2015-2020.

J. Epstein. *Generative Social Science: Studies in Agent Based Computational Modeling*. Princeton Press, 2006.

M.C. González, C.A. Hidalgo and A.-L. Barabási. Understanding individual human mobility patterns. *Nature* 453 (2008), 779-782.

N. Gilbert and K. Troitzsch. *Simulation for the Social Scientist*. Open University Press, Philadelphia, 1999.

D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the Spread of Influence in a Social Network. In *Proc. KDD*. 2003.

R. Little. Controlling Cascading Failure: Understanding the Vulnerabilities of Interconnected Infrastructures. *Journal of Urban Technology* 9, 1 (2002), 109-123.

H.S. Mortveit and C.M. Reidys. *An Introduction to Sequential Dynamical Systems*. Universitext Series, Springer Verlag, 2007.

M. Macy and R. Willer. From Factors to Actors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology* 28 (2002), 143-166.

M. Newman. The structure and function of complex networks. *SIAM Review* 4 (2003).

Critical Infrastructure Protection and the Law: An Overview of Key Issues. National Research Council of the National Academies National Academic Press, Washington D.C.

T. Sandler. *Collective action: Theory and applications*. U. Michigan Press, 1992.

Biographies

Chris Barrett, Ph.D. is Professor of Computer Science, Director of the Network Dynamics and Simulation Science Laboratory, and Executive Director of the Biocomplexity Institute of Virginia Tech.

Jeffrey Johnson, Ph.D. is Professor of Complexity and Design in the Faculty of Science, Technology, Engineering and Mathematics, The Open University.

Madhav Marathe, Ph.D. is Professor of Computer Science in the Network Dynamics and Simulation Science Laboratory, Biocomplexity Institute of Virginia Tech.

DOI: 10.1145/3158342