Information Integration to Support Model-Based Policy Informatics

Christopher L. Barrett, Stephen Eubank, Achla Marathe, Madhav V. Marathe, Zhengzheng Pan, Samarth Swarup

Network Dynamics and Simulation Science Laboratory,
Virginia Bioinformatics Institute,
Virginia Tech,
Blacksburg, Virginia 24061.
swarup@vbi.vt.edu

I**nformation Integration to Support Model-Based Policy Informatics**

Christopher L. Barrett, Stephen Eubank, Achla Marathe, Madhav V. Marathe, Zhengzheng Pan, and Samarth Swarup

ABSTRACT

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. We describe our approach to this problem, and how we are building 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. Our simulation-based approach 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. We present the rationale and design of our methodology and discuss several areas of actual and potential application.

Keywords: Policy Informatics, Information Integration, Public Policy, Complex Systems, Computational Social Science

The Challenge

Policy planners often look for quick answers to "what if" questions: What would happen if this cell tower became non-operational? What would happen if the people in one demographic group were vaccinated? The fact is, the answers to such questions are never quite quick or simple.

Many public policy issues nowadays involve Biological, Informational, Sociological, and Technological (BIST) systems, which consist of a large number of interacting physical, biological, and human/societal components whose global system properties are a result of interactions among lo- cal system elements (Albert and Barabási, 2002; Eubank, Guclu, Kumar, Marathe, Srinivasan, Toroczkai, and Wang, 2004; Newman, 2003; Vega-Redondo, 2007). In other words, the behaviors of each component and the interactions among groups of components have an effect on the outcome at the system level as well as the global level. At the same time, elements' behavior and interactions are affected by the global state. The interdependencies between the elements' behavior and effects on the global outcome show that it is a two-way feedback process which makes it difficult to control these systems (Sterman, 2006). Also, they involve multiple stakeholders, who often have conflicting optimization criteria.

These interactions and interdependencies can be abstracted representationally as networks, and network science provides 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 (Newman, 2003; Eubank et al., 2004; Albert and Barabási, 2002; Barabási and Albert, 1999; Barrett, Eubank, Kumar, and Marathe, 2004; Barrett, Eubank, and Marathe, 2006).

Note that different types of interactions are carried out on different networks. For instance, diffusion of knowledge, news, and rumors take place on information networks including media coverage and word-of-mouth interpersonal channels. Physical cascades occur on infrastructural networks, e.g., an evacuation may result in traffic congestion as well as a breakdown of the communication system through congestion on the base station (Barrett, Beckman, Channakeshava, Huang, Kumar, Marathe, Marathe, and Pei, 2010).

Networks are not static but evolve over time. In case of social networks, the change reflects both day-to-day randomness in contacts and systematic changes due to behavior adaptation (Chen, Marathe, and Marathe, 2010; Newman, 2003; Vega-Redondo, 2006; Young, 1998) For instance, physical networks such as transportation networks change due to road closures, new bridges, high- ways etc., while informational networks like mobile ad hoc networks are self-configuring networks of mobile devices which change with the movement of mobile devices (Barrett et al., 2010; Atkins, Chen, Kumar, and Marathe, 2009).

Since networks are often closely correlated, we have not only evolving networks, but co- evolving networks. For example, consider flu control and prevention. In flu season, there are public level health controls, as well as privately imposed self-interventions. The specific method may be pharmaceutical, such as vaccination or anti-viral treatment, or non-pharmaceutical such as social distancing (not go to work, close schools). Each method leads to a co-evolution of epidemics, behavior, and networks.

The flu virus is transmitted through social contacts. Therefore to model the spread of the disease in detail, one needs the social contact network of the entire population. Changes in the structure of this network are coupled with the health state of each individual. An intervention, whether pharmaceutical or not, changes the epidemic dynamics through changes in the contact network or changes in the probability of infection.

These changes occur through changes in individual behavior in response to the epidemic dynamics. For instance, in case of a big outbreak with high infection count, people are more likely to get vaccinated. Such decisions are typically based on information received through a person's social and information network, and in turn result in changes in the contact network and the epidemic dynamics.

The co-evolution of network structures and the local interactions in each network are often results of 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 as well. For instance, the networks considered are often either non-stochastic or from random graph family in order to keep the analysis tractable (Albert and Barabási, 2002). Individual characterization is either very limited or non-existent (Jackson, 2007; Jackson and Yariv, 2009).

Since policy problems are often trans-disciplinary, adequate solutions require integration of models from multiple fields and data from multiple sources. Integration of information is a particularly challenging problem, especially in the face of massive data sources that have been collected by different individuals and institutions in parallel, and rarely specifically for the issue of interest. Consequently, such integration requires rigorous statistical methods with deep understanding of the dynamics and complexity of the systems involved.

Networks provide a representation capable of integrating these multiple data sources, and their multiplex interactions. Beyond integrating the information, using it for policy and decision-making presents its own challenges of scale. For example, to study the spread of epidemics in New York City, a network with 17 million nodes (individuals) and about a billion edges (contacts) is required. For the entire nation, it would be 300 million nodes and 22 billion edges. With this kind of scale and computational needs, high-performance computational tools and efficient algorithms are essential (Bisset and Marathe, 2009).

Hence 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 allowing evaluation of a large class of possible interventions (flexibility). The team at the Network Dynamics and Simulation Science Laboratory (NDSSL) at Virginia Bioinformatics Institute (VBI) at Virginia Tech has developed a methodology to integrate information from multiple sources and to build large-scale high-resolution simulations to address these challenges (Barrett, Bisset, Eubank, Feng, and Marathe, 2008). Next, we describe the rationale and design of our approach and how it satisfies the four properties mentioned above.

Information Integration and Simulation

Digital data are being generated at an amazing rate. It is said that if we add up all the digital data generated in a year, through computers, mobile phones, digital cameras, television, etc., we are entering the yottabyte[1] 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 that we do not have the right data for the problem at hand.

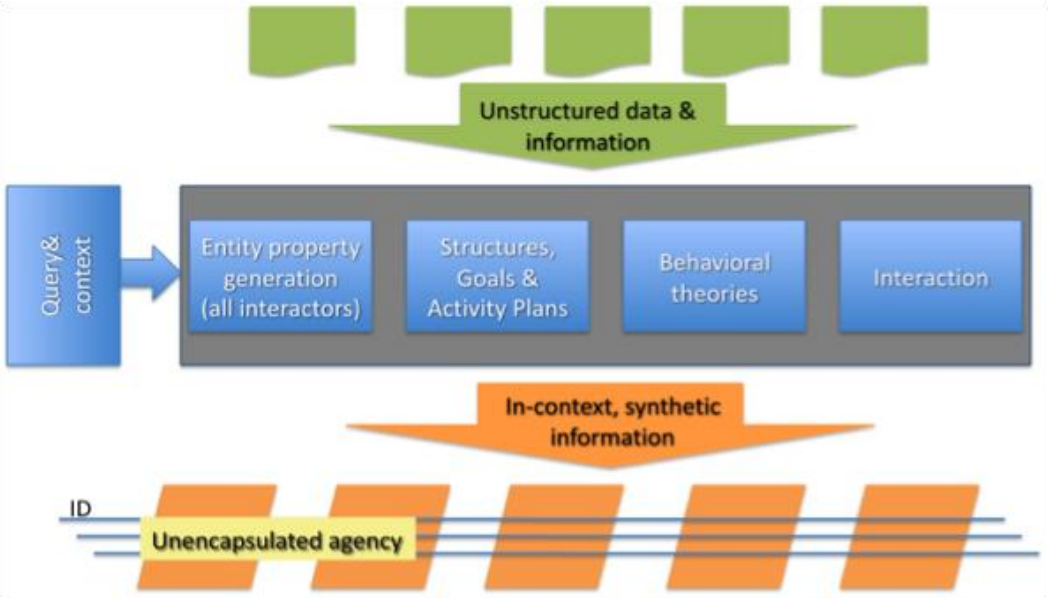
Available observations and knowledge are normally not structured specifically for a particular question. We overcome this data problem by using available, sometimes imperfect information in the form of data and procedures, to synthesize an integrated representation of what is known in the context of the decision to be made.

In our general framework, we use multiple modules that serve different functions. The most important characteristic of the framework is the control of data flow. Namely, we have multiple (external) data sources as the inputs that are used by and separated from, the modules. One particular data set could be used by multiple modules. The (output) data created by modules are also used as inputs to other modules. The separation and flow of data is the key as it allows the information integration process to perform iterative refinement, which is critical because that enables the specifics of the integrated information to be determined by the context of the problem. For instance if we are interested in studying the spread of a disease, getting the traffic patterns exactly right might not be necessary, but if we are studying the transportation infrastructure, then we need to include data on traffic patterns, road capacities, speed limits etc. into the information integration process. Figure 1 below shows a simplified overview of such framework.

From an informatics perspective, there are a couple of important things to note about this approach. First, it goes beyond traditional informatic 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. Second, this is really a "generative social science" approach (e.g., Epstein, 2005), since the generated model can be more complete than information gained from surveying alone.

[1] 1 yottabyte = 10^{24} bytes = 1 trillion terabytes.

Figure 1: Overview of the information integration framework



Additionally, the synthetic data created (population, contact networks, activities, etc.) by our approach based has the following features: (1) it is statistically equivalent to the real data, (2) it is anonymous which helps overcome issues related to human subjects, (3) it is comprehensive and provides justification for certain kinds of policy decisions, (4) whenever available, components of synthetic data can be replaced with real data, and (5) it represents integrated interaction-related information from multiple sources.

A complete system to support policy and decision-making involves not just the integration of information and data from multiple sources, but also the software and high-performance-computing infrastructure required to conduct large-scale simulations. This is a very important issue because experimental designs in these domains generally involve multiple factors and many iterations to bound the variance. It is necessary to be able to run complete experiments on the time-scale of hours in order to help guide policy in domains like epidemiology, where decision makers only have hours or days to evaluate interventions in the midst of an outbreak.

To solve this issue, two software systems have been developed and implemented at NDSSL, Simdemics (Bisset and Marathe, 2009; Bisset, Feng, Marathe, and Yardi, 2009a; Barrett et al., 2008) and EpiFast (Bisset, Chen, Feng, Kumar, and Marathe, 2009b). Both are capable of efficiently simulating various contagion processes on large-scale distributed-memory computing infrastructure. Simdemics works with the full person-location dynamic synthetic population, and has a very expressive specification system for defining interventions and the resulting alterations to people's daily schedules and interactions. EpiFast works with the person-person social contact network, or any other network where every node is of the same type, and does a very rapid evaluation of a diffusion process on the network. They both implement highly-parallel algorithms that automatically distribute the network over the available processing elements and manage the communication between elements to keep the state of the diffusion process consistent.

The final component required to develop simulations in the epidemiological domain is a model of the contagion itself. For example, in epidemic modeling, each disease has specific parameters such as incubation time, probability of infection, etc. Also, different models of propagation can be used, such as SIR (Susceptible-Infectious-Recovered) or SEIR (Susceptible-Exposed-Infectious- Recovered). Our tools allow user-defined configuration to accurately specify these parameters. Similarly, social contagion processes, like the spread of smoking behavior, can involve other factors that determine the probability of spreading, such as popularity, socio-economic background, age, prior exposure, and many others. In such cases, we develop data-driven models of diffusion based on domain-specific data sets (e.g., Harris, 2008). We now discuss how the use of integrated information allows one to build systems with the properties required for a systematic solution to policy problems.

Adaptability

Most policy problems are trans-disciplinary. For example, a disease epidemic is not just a public health problem, it is also a social and economic problem. Epidemics place a huge cost upon society. It is estimated that the 1918 flu pandemic resulted in about 50 million deaths worldwide, and that a similar pandemic today would result in 150 million deaths and cost $4.4 trillion. These costs come from loss of income and productivity, interventions, distribution of vaccines and anti-virals, school closures, and caring for sick and children. Consequently, an effective tool for policy informatics needs to be a multi-theory, multi-actor, multi-perspective system. More concisely, we refer to this property as adaptability.

Modeling such a complex system requires a multi-theory and multi-perspective approach. Al- though researchers have been attempting to use general systems theory as a unified theory to model system problems across fields since the work of Von Bertalanffy (1968), major weaknesses of this approach have yet to be overcome (e.g., Kast and Rosenzweig, 1981). Besides, it is necessary to investigate complex issues with perspectives from different fields, as no single theory can explain all the aspects of issues that are observed (Contractor, Wasserman, and Faust, 2006). For instance, developing an integrated epidemic model requires input from multiple theories, including biology (e.g., to parameterize models for specific diseases), economics (how many doses of anti-virals to produce and at what price), sociology (whom to vaccinate) and public policy (which interventions to apply and when), among others.

From the analysis point of view, the system supports multiple perspectives. Designing interventions and vaccination policies requires multi-criterion optimization because individual objectives may not be socially optimal. For example a public policy of complete school closure or mass vaccination of all individuals may be unacceptable or unimplementable. Examples of the use of our system in multi-perspective analysis are given in section 3.

As mentioned above, the integration of models is more than a simple mix. Oreskes (2000, 2003) and Oreskes and Belitz (2001) have described some of the fundamental issues with building complex models. The basic issue is a seemingly inverse relationship between complexity and realism of the model on the one hand, and trust in the model and certainty in its output on the other. As we include more and more factors within a model, the uncertainty in its

output also increases. Model integration also goes beyond model federation, which has typically been driven by the desire for point estimates. However, for the kinds of complex systems in which we are interested, this is simply unachievable.

Third, and equally important, effective policy in this domain requires input from multiple stake- holders. For example, government and other institutions at multiple levels are involved in deciding and implementing intervention policies, awareness campaigns, etc. To this end, we have designed a web-based, service oriented front-end to the simulation environment, called DIDACTIC, which allows epidemiologists, policy planners and researchers to use our framework. This has the advantage of making the tools directly available to the stakeholders, which in turn allows the computational modeling aspect to be a real participant in the policy conversation. Policy discussions can lead directly to simulation experiments, and the simulation results can be directly accessed and interpreted by the policy planners. This also has the secondary advantage of providing a sense of participation and ownership to the stakeholders. The simulation environment is no longer just a black box to them, since they are trained on how it works, and can design experiments and analyze results themselves. The system then takes care of translating the experiment design into a set of compute jobs that run on the HPC infrastructure and delivers the results back to the user.

To facilitate the specification and implementation of interventions, a database tool called Indemics has been designed and developed. Indemics is an Interactive Epidemic Simulation and Modeling Environment that allows a user to actively interact with the system so that the user can make changes to the social network, individual behaviors, or the disease models in run time (Bisset, Chen, Feng, Ma, and Marathe, 2010). It supports rich queries across multiple data types, e.g., find a count of infected persons in zip code 24060 or find all the infectious students in Blacksburg High School and their family members. The user interacts with Indemics using well-defined languages, e.g., count infected persons: group = seniors, infected day = between 20 and 22. One can also build pre-defined libraries of queries by expert users.

The software infrastructure also includes a digital library which will keep archives of the old simulation runs, disease models, configurations, input and output files (Leidig, Fox, Marathe, and Mortveit, 2010). Our graph analysis software library Galib provides efficient implementations of various classical and new graph measures that are motivated by the analysis of social contact graphs and disease dynamics on such graphs. It can be used to compute efficiently structural measures of social contact networks with 10 million vertices and over 500 million edges.

In addition, adaptability also means that the system must be able to respond to a dynamic world, by providing means for continual data uptake, state assessment, decision analysis, and action assignment. For example, during an epidemic outbreak, new data are constantly being collected on the status of the outbreak, such as number of people infected, their demographics, their locations, etc. Our simulation system allows continual (or periodical) updating of the state of the model based on the new data runs essentially like a predictor-corrector filter – the model generates a prediction of the state of the world, and

input from the world is used to correct this prediction. The prediction error then adjusts the model so that subsequent predictions are closer to the mark. Indemics allows our simulation system to be run in this interactive way, which allows the model to be more veridical. For example, each day, the state of the model can be adjusted to reflect the number of people infected, as determined by the data gathered by the CDC. Thus the decisions based on the model always reflect the latest information.

Extensibility

As discussed earlier, we integrate information from multiple sources to design our modeling and simulation system. In principle, we can keep adding new data sources and keep expanding the integrated information being generated[2]. Thus, our models are extensible by design. However, doing this in a coherent manner requires some understanding of the data. There are four different kinds of data sources: survey data, administrative data, commercial data and increasingly, ubiquitous in- formation sources like the Internet. Each has different properties and present different challenges to integration.

Survey data are typically expensive to gather, but are nevertheless highly desirable because they are gathered in a very controlled and rigorous manner which means we have a good understanding of the sources of error and variance in the data. These are the data sources that have traditionally been used for social theories and therefore can provide a sound theoretical grounding for a model. However, survey data are always sparse, and always slightly outdated. For instance, the census is only conducted every ten years.

Administrative data on the other hand, are gathered by bureaucratic institutions such as the Department of Motor Vehicles and the Internal Revenue Service, or by corporations and institutions such as hospitals, universities, etc. These data tend to be the opposite of survey data in most respects: they are much more complete and current, but much less rigorously gathered, often containing errors, missing values, and unknown biases and sources of variance.

Commercial data are the data that are commercially available from companies such as Dun and Bradstreet, Acxiom, Navteq etc. These data are expensive to obtain and keep current.

Ubiquitous information sources, such as the Internet, cell phone data, etc., are even more un- controlled, but are current and dynamic. However, it is not clear as to how to harvest these datasets and how to integrate and reconcile them with the other kinds of data sources.

Much of the research in information integration lies in developing techniques to fuse various sources of data in statistically rigorous ways, and in recognizing gaps in the data and developing well-founded models to fill in these gaps. For example, to model the diffusion of smoking, we need adolescent friendship networks for the population of each school in the synthetic population. For this we have adapted and generalized a method of hierarchical network decomposition due to Clauset, Moore, and Newman (2008). They showed how to generate a dendrogram that represents the hierarchical clustering structure present in a network.

[2] Note that extensibility refers to addition of data sources of a new kind, such as adding consumer behavior survey data to the census data. Whereas updating data mentioned in adaptability refers to new data added to the existing data source, such as the latest count of infected individuals.

We applied their technique to all the adolescent friendship networks in the Add Health data (Harris, 2008), and then developed a generative probabilistic model over the resulting dendrograms. This probabilistic model then allows us to generate new dendrograms of different sizes, and thereby to generate friendship networks for our synthetic adolescent populations. Problems like these spur the development of new techniques in network science, machine learning, and related fields.

Scalability

Realistic simulations require the ability to compute interactions between millions of agents. This scalability problem has been referred to earlier, in the context of the need for high-performance computing infrastructure and highly parallelized algorithms. However, the main insight behind our models that allows scalability is the notion of interaction-based computing (Barrett et al., 2006). This concept is best explained through an example.

Consider the problem of representing traffic dynamics in a city. There is no explicit algorithmic description of this problem. Traffic is an emergent property from the interactions between individual drivers. It is possible to make a careful study of individual drivers to determine a detailed model of their behavior and responses to various conditions and situations, and thereby to make a model that allows simulation of traffic dynamics. In most agent-based models that follow this approach, each agent would carry a complete set of driving rules, or possibly some other statistical representation. Such an approach results in "heavy" agents, i.e., the description of each agent is complex and memory-intensive, which hinders scaling.

However, it turns out that an alternative model is possible, which uses a simple and parameterized description of the individual driver, but still results in the correct traffic dynamics because they emerge from the interaction pattern between agents. In this case, the computing that was being done inside the agent in the previous model is now being done through the interaction between agents and their neighbors. Moreover, this interaction is dynamic and the neighborhood changes all the time. In other words, the environment is not static. The driver interacts continually with the environment and co-evolves with it. For instance, Barrett, Wolinsky, and Olesen (1996) showed that a simple set of cellular automaton rules for traffic simulation can be viewed as equivalent to an adaptively compensated derivative feedback control system. This macro-scale equivalence and parameterized specifications for each individual allows agents to be much simpler in description, i.e., they are "light-weight", which greatly enhances the scaling capability (Atkins, Barrett, Beckman, Bisset, Chen, Eubank, Feng, Feng, Harris, Lewis, Kumar, Marathe, Marathe, Mortveit, and Stretz, 2008).

In our simulation models, agents are therefore properly viewed as "unencapsulated". In other words, the description of the individual agent is distributed across the entire integrated information data structure, rather than being contained in a single software object. This prevents the problem of over consuming memory by using only the information needed for each agent and interaction. In case more details and data are needed, users can turn to the corresponding module.

Flexibility

One of the main advantages of a simulation-based approach over a more traditional differential equation based approach to policy modeling is the ease and intuitiveness with which interventions can be represented. For example, in the epidemiological domain, the notion of "social distancing" is an important basis for designing interventions. The basic idea is that by reducing the number of edges in the social contact graph, we can reduce the available paths for the spread of disease, thereby reducing the size of the outbreak. Differential equation based models typically result in recommendations such as "reduce the number of social contacts by 20%". These can be further qualified by demographic data, but the essential recommendation remains the same. However, it is unclear how to implement such a recommendation in the real world. Should the schools remove 20% of the students from attendance, should each household member reduce contact with other members by 20%, should grocery stores allow 20% less customers, etc.

A simulation-based approach, on the other hand, allows policy makers to experiment with very concrete and specific interventions, such as closing particular schools for particular periods of time. Note that the intervention is not to close all schools, or to close some randomly chosen schools, but to close a specific set of schools. Further, the simulation also allows us to accommodate resulting changes in activity schedules for parents, children and teachers, which further determines the revised social contact network. A large class of interventions has been implemented within Simdemics, including vaccination, anti-viral distribution, and school closures, generic social distancing, household quarantine etc. It is also easy to mix and match these interventions, so that multiple ones can be applied at the same time.

Internally, each agent in the simulation is represented by a probabilistic timed transition system (PTTS). This can be thought of as a set of finite-state automata, where the transitions in each automaton can be triggered by the states of the other automata, the states of neighbors' automata, or time (e.g., to represent a transition from infectious to recovered state). This means that although the automata are represented separately, the system is effectively operating in the cross-product space of these automata. This results in a very flexible and scalable representation, which is powerful enough to represent not just disease dynamics but also social contagions.

Practical Applications

Our modeling environment has been used in a number of user-defined case studies including re- cent pandemic planning studies undertaken for DoD and DHHS. Multiple studies were conducted for the DoD regarding military preparedness and force readiness. The studies elucidated how protecting a small critical subset of a larger population is fundamentally different from public health epidemiology. The studies provided guidelines for military preparedness in the event of an epidemic outbreak. The results showed the importance of early detection in implementing effective sequestration and the apparently counter-intuitive result that sequestration, if implemented late, might lead to more infections rather than less infections (Atkins, Barrett, Beckman, Bisset, Chen, Eubank, Lewis, Marathe, Marathe,

Mortveit, Stretz, and Kumar, 2006b; Atkins, Barrett, Beckman, Bisset, Chen, Eubank, Kumar, Lewis, Macauley, Marathe, Marathe, Mortveit, and Stretz, 2006a). These studies have guided the continued evolution of our simulation system both in terms of its usability and model development. The studies also helped us identify new research questions at the interface of multi-agent modeling, data mining, network science and high performance computing.

Next we describe a particular case study that demonstrates the applicability of our modeling approach to real life scenarios (Chen et al., 2010). This is the first study that uses individual based approach to analyze how behavioral changes occur in response to the growth of the disease and how these changes, in turn, affect the disease dynamics. In order to study the diffusion of disease on social networks, we first build a social network.

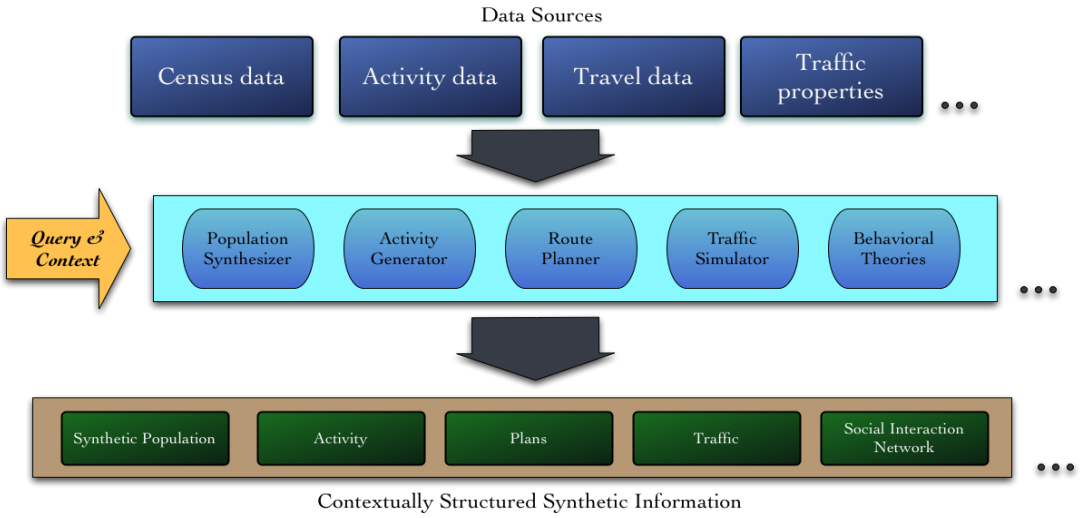
Information Integration for Building Social Networks

The key input in this study is the synthetic population and social contact network on which the spread of flu virus occurs. To generate the synthetic populations and the social contact network, a multi-step process is involved (Bisset and Marathe, 2009; Beckman, Baggerly, and McKay, 1996)[3]. The process is illustrated in Figure 2.

First a synthetic set of individuals for a particular US area is built by integrating a variety of commercial and public data sources including the US Census. The synthetic population is a set of synthetic people and households, located geographically, each associated with a set of demographic variables drawn from the Census. It is generated in such a way that a census of the synthetic population is statistically indistinguishable from the original census, i.e., the joint distributions of demographics (such as age, gender, and income) are fitted to match those taken from the Census (Barrett, Beckman, Berkbigler, Bisset, Bush, Campbell, Eubank, Henson, Hurford, Kubicek, Marathe, Romero, Smith, Smith, Speckman, Stretz, Thayer, Eeckhout, and Williams, 2001; Beckman et al., 1996).

[3] For more details on algorithms used in the process and the synthetic data, please refer to the technical reports available at http://ndssl.vbi.vt.edu/transims.php and http://ndssl.vbi.vt.edu/opendata/index.php.

Figure 2: Schematic of the information integration process



The Census provides data aggregated to the block group level, essentially providing marginal distributions and covariances for a number of demographic variables. From these, we re-construct a disaggregated population using a method called iterative proportional fitting (Beckman et al., 1996). The resulting synthetic population matches the statistics of the census data at the block group level, but not at a finer granularity (e.g., not at the level of individuals). In this sense, a synthetic population is inherently anonymous since it cannot be used to identify particular individuals in the real world. This is an important issue in many policy domains.

We then assign a set of activities to each person in the household based on activity-time survey data. A set of activity templates for households is determined based on several thousand responses to an activity or time-use survey (Barrett et al., 2001; Beckman et al., 1996). The activity templates include the sorts of activities each household member performs and the time of day they are performed. Each synthetic household is then matched with one of the survey households using a 12-parameter decision tree. Based on demographics such as the number of workers in the house- hold, number of children, their ages, and the characteristics of a place such as size and distance between the place and an individual's home location, we measure the likelihood that an activity happens at a specific location. In turn, the synthetic household is assigned the activity template of its matching survey household. For each household and each activity performed by this household, a preliminary assignment of a location is made based on observed land-use patterns, tax data, etc.

A complete daily schedule for each individual gives a bipartite person-location network, where nodes are persons or locations, and a link between a person and a location exists when that person visits that location. The link is labeled with the start and end time of the visit, which effectively means that the graph is time-varying. From this graph, we can also induce a time-varying person- person interaction graph, by adding a link between each pair of persons who are at the same location for an overlapping duration. Then we can remove the location nodes, and we are left with a time- varying social contact network.

In other words, we derive a person-person social contact network from the person-location activity network. Note that we have extracted this network from multiple sources of data. It is not directly available. Given the start and end time of each person's stay at a location, we also have the contact duration for each pair of persons. For epidemiology studies, the edges of the social contact networks are weighted with the conditional transmission probabilities based on the duration of contact.

Case Study

This case study models the spread of a flu-like illness in the New River Valley region of Virginia with a population of 150,000. We assume that there are 15,000 units of anti-viral courses available. The anti-virals can be distributed through two channels: the public sector and the private sector. The goal of this research was to find an optimal distribution of a limited supply of anti-virals between the public sector and the private sector so that the attack rate is minimized and enough revenue is generated to recover the cost of the anti-virals (Chen et al., 2010).

The public sector distribution of anti-virals is done through the hospitals. At a hospital, if an individual is diagnosed to be infected, s/he is given the anti-viral at no cost. This study accounts for the fact that infected individuals do not always show symptoms, and symptomatic individuals do not always report to hospitals. Also, misdiagnosis of sick and worried-well is possible.

The private sector distributes the anti-virals through the market where individuals can purchase the anti-virals for prophylactic use or future treatment. The revenue from the market helps recover the overall cost of anti-virals. The private demand for anti-viral is based on the budget of the household, price of the anti-viral, number of infections in the society and the demand elasticity of prevalence.

In addition to buying anti-virals, the household members isolate themselves at home when a member of the household is diagnosed to be infected. These interventions change the epidemic dynamics through changes in the social contact network and transmission probabilities. The change in epidemic changes the disease prevalence, which affects the private demand for anti-virals, which in turn impacts the health state of the individuals (Barrett, Bisset, Chen, Lewis, Eubank, Kumar, Marathe, and Mortveit, 2007).

The simulation results show that allocating the entire stockpile of the anti-virals to just the public sector or just the private sector is a sub-optimal strategy. The study isolates the effects of changes in behavior and changes in the social network caused by people's reaction to disease prevalence, on the prevalence itself. The results showed that prevalence elastic demand of anti- virals can delay the onset of the outbreak and changes in the social network caused by quarantine can reduce the peak of the epidemic curve by a significant amount.

The attack rate decreases as more of the anti-viral stockpile is allocated to the hospitals, since it is targeted to those who are infected. However, the attack rate reaches a lower bound because only a fraction of the infected individuals report themselves to the hospitals and get correctly diagnosed.

Another finding is that the market stockpile is taken up by people according to their household income rather than their health state. In other words, households with high income get most of the private anti-viral stockpile even though the exposure count among them is low. This is more so when the demand function is prevalence elastic. In this case, the households in the top 30 percentile of income get the entire private supply of anti-virals. This is because in case of elastic demand, when disease prevalence is high, the demand for anti-viral goes up which drives the price higher, making it unaffordable to lower income families.

In summary, this case study examines the co-evolution of epidemics, individual behavior, and social networks. The simulation results show that the dependence of demand on disease prevalence postpones the peak of the epidemic by about a month; and household isolation decreases the peak attack rate by more than 1000.

Conclusions and Outlook

We are living in a world of networks, and these networks are becoming more interdependent every day. Social networks are inextricably entangled with communication networks, transportation networks, logistical networks, and the like. The increasing dependencies and the increasing densities of these networks imply that a disruption in any one affects all the others. For a policy to be effective in such complex systems, response has to be very rapid, but their very complexity makes decision-making difficult and time-consuming.

The challenge, therefore, is to confront this complexity and to build the tools to tame it, so that policy-makers can leverage from the best models and computing capabilities available for decision-making. Our vision is to develop a computational environment for policy informatics that can seamlessly integrate many sources of data with the high-performance computing hardware and software necessary to process them, so that they can be delivered to a policy decision-maker as cleanly as the Google home page. This is not a static system. As the underlying theory of network science evolves and the underlying computing technology evolves, our tools will evolve as well, but much of this can be transparent to the end-user.

Validation of integrated information is a new challenge for information science, because classical techniques are insufficient. Techniques for validation of each data source and survey separately exist, and are well understood in statistical science and survey science (e.g., Rice, 2006), but it is still a developing science in the field of information integration.

For the past sixteen years, the team members of the Network Dynamics and Simulation Science Laboratory (NDSSL) have been pursuing research on biological, informational, social, and technological networks to further this vision of integrated situational awareness and consequence analysis. Such an environment can support policy-makers from the highest levels to the levels of first responders.

In this article we have described the philosophy and design of our approach to information integration and simulation for policy informatics. Our systems are built to support

multiple policy domains, from multiple perspectives. We have discussed how they compare with older approaches to systems modeling and agent-based simulation. We have also presented a specific case study that demonstrates their use. We believe that the future of policy and decision-making is data-driven. Our work aims at integrating data with models and procedural knowledge to create the necessary systems to support the accelerating pace and complexity of human social and cultural life.

About the Authors:

Dr. Christopher L. Barrett is the Director of Network Dynamics and Simulation Science Laboratory of the Virginia Bioinformatics Institute at Virginia Tech and Professor of Computer Science. His work includes the development of large-scale, high performance simulation systems, distributed computing for simulation-based study of large socio-technical systems and formal approaches to interaction-based systems. Before moving to Virginia Tech in 2004, he led the Basic and Applied Simulation Science Group at Los Alamos National Laboratory. Dr. Barrett and his group have developed many large simulation environments, including TRANSIMS, EPISIMS, distributed communication and sensor systems, Marketecture, Urban Infrastructure Suite and co- established the National Infrastructure Simulation and Analysis Center (NISAC) at DHS. He is PI/Co-PI on several DTRA, NIH, and NSF projects related to national and international crisis planning and response, and developing novel multi-theory agent-based HPC models coupled with analytical and algorithmic techniques for inference and state assessment of complex socio-technical networks.

Dr. Stephen Eubank is a research professor at the Virginia Bioinformatics Institute at Virginia Tech, where he serves as Deputy Director of the Network Dynamics and Simulation Science Laboratory. He received his B.A. in physics from Swarthmore College and his Ph.D. in theoretical physics from the University of Texas at Austin. He has worked in the fields of financial market modeling (as co-founder of Prediction Company); ecological time series analysis (at Biosphere 2); natural language processing (at Advanced Telecommunication Research in Kyoto); fluid turbulence (at the La Jolla Institute); nonlinear dynamics, chaos, and simulation of socio-technical infrastructure (at Los Alamos National Laboratory); and mathematical epidemiology (as a PI within the NIH/NIGMS Modeling Infectious Disease Agent Study network). He is keenly interested in the fundamental problem of understanding complex systems wherever in society they are found: what macroscopic phenomena result from microscopic interactions in a complex interaction network topology?

Dr. Madhav Marathe is a Professor, of Computer Science, and Deputy Director of Network Dynamics and Simulation Science Laboratory (NDSSL), Virginia Bioinformatics Institute at Virginia Polytechnic Institute and State University (VT). He leads the basic and applied research program in modeling and simulations of large complex networks and the development of associated computational technologies to support this program. Before coming to Virginia Tech, he was a Team Leader in the Basic and Applied Simulation Science group 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 VT, he and NDSSL team members are working on a number of projects funded by DTRA, NSF, NIH, CDC, and DARPA, related to human and social dynamics, computational epidemiology and public health, network science, communication networks, critical infrastructure protection and high performance computing.

Dr. Achla Marathe is an associate professor at the Agricultural and Applied Economics Department at Virginia Tech. She is also the lead economist and social scientist at the Network Dynamics and Simulation Science Laboratory at the Virginia Bioinformatics Institute at Virginia Tech. Before coming to Virginia Tech in 2005 she worked at the Los Alamos National Laboratory for 10 years on topics varying from fraud detection in health care, to modeling and simulation of restructured electricity markets. She has been interested in issues that are at the intersection of public health, economics and individual behavior. She has also been developing detailed spatio-temporal models of networked markets such as electricity and wireless spectrum using activity-based synthetic social networks.

Dr. Zhengzheng Pan is a postdoctoral associate at the Network Dynamics and Simulation Science Laboratory at the Virginia Bioinformatics Institute at Virginia Tech. She earned her B.S. in Computer Science and Technology from Chu Kochen Honors College of Zhejiang University and her Ph.D. in Economics from Virginia Tech. Her primary research interests are microeconomic behavior, interaction, and decision-making in the context of networks. Specific topics include social learning, game theory, and cost-benefit based network formation. She now works with a multi-disciplinary team of scientists and researchers to develop theoretical foundations of very large socio-technical and information networks, as well as stakeholder-defined case studies for in- forming public policy in various areas such as public health, market regulation, and infrastructure networks.

Dr. Samarth Swarup is a postdoctoral researcher at the Network Dynamics and Simulation Science Laboratory at the Virginia Bioinformatics Institute at Virginia Tech. Before coming to Virginia Tech in 2008, he earned his Ph.D. In Computer Science from the University of Illinois at Urbana-Champaign and spent a year as a postdoctoral researcher at the Graduate School of Library and Information Science at UIUC. His research interests lie in computational social science, computational sociolinguistics, artificial intelligence, machine learning, and language evolution.

Acknowledgment

We thank our external collaborators and members of the Network Dynamics and Simulation Science Laboratory (NDSSL) for their suggestions and comments. This work has been partially supported by NSF Nets Grant CNS- 0626964, NSF HSD Grant SES-0729441, NIH MIDAS project 2U01GM070694-7, NSF PetaApps Grant OCI-0904844, DTRA R&D Grant HDTRA1-0901-0017, DTRA CNIMS Grant HDTRA1-07-C-0113, NSF NETS CNS-0831633, DHS 4112-31805, DOE DE-SC0003957, NSF REU Supplement CNS-0845700, US Naval Surface Warfare Center N00178-09-D-3017 DEL ORDER 13, NSF Netse CNS-1011769 and NSF SDCI OCI-1032677.

References

Albert, R., and A. Barabási. 2002. Statistical mechanics of complex networks. *Review of Modern Physics* 74:47–97.

Atkins, K., C. Barrett, R. Beckman, K. Bisset, J. Chen, S. Eubank, A. Feng, X. Feng, S. Harris, B. Lewis, V. S. A. Kumar, M. V. Marathe, A. Marathe, H. Mortveit, and P. Stretz. 2008. An Inter- action Based Composable Architecture for Building Scalable Models of Large Social Biological, Information and Technical Systems. *CTWatch Quarterly* 4(1).

Atkins, K., J. Chen, V. S. A. Kumar, and A. Marathe. 2009. Structure of electrical networks: A graph theory based analysis. Invited to a special issue in the *International Journal on Critical Infrastructure* 5(3): 265–284.

Atkins, K., C. Barrett, R. Beckman, K. Bisset, J. Chen, S. Eubank, V. S. A. Kumar, B. Lewis, M. Macauley, A. Marathe, M. Marathe, H. Mortveit, and P. Stretz. 2006a. Simulated pandemic influenza outbreaks in Chicago: NIH DHHS study final report. Tech. Rep. 06-023, NDSSL.

Atkins, K., C. Barrett, R. Beckman, K. Bisset, J. Chen, S. Eubank, B. Lewis, A. Marathe, M. Marathe, H. Mortveit, P. Stretz, and V. S. A. Kumar. 2006b. DTRA case study to sequester critical subpopulations. Tech. Rep. 06-060, NDSSL.

Barabási, A. L., and R. Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286(5439): 509.

Barrett, C., R. Beckman, K. Berkbigler, 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. 2001. TRANSIMS: Transportation analysis simulation system. Tech. Rep. LA-UR-00-1725. An earlier version appears as a 7 part technical report series LA-UR-99-1658 and LA-UR-99-2574 to LA-UR-99-2580, Los Alamos National Laboratory Unclassified Report.

Barrett, C., K. Bisset, J. Chen, B. Lewis, S. Eubank, V. S. A. Kumar, M. Marathe, and H. Mortveit. 2007. Effect of public policies and individual behavior on the co-evolution of social networks and infectious disease dynamics. In *Proceedings of the DIMACS/DyDAn Workshop on Computational Methods for Dynamic Interaction Networks*.

Barrett, C., K. Bisset, S. Eubank, X. Feng, and M. Marathe. 2008. EpiSimdemics: An efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *Proc. of the ACM/IEEE conference on Supercomputing* (SC), 1–12. Piscataway, NJ, USA: IEEE Press.

Barrett, C., R. Beckman, K. Channakeshava, F. Huang, V. S. A. Kumar, A. Marathe, M. Marathe, and G. Pei. 2010. Cascading failures in multiple infrastructures: From transportation to communication network. In The Fifth International CRIS Conference on Critical Infrastructures. Beijing.

Barrett, C., S. Eubank, V. S. A. Kumar, and M. Marathe. 2004. Understanding large-scale social and infrastructure networks: A simulation-based approach. *SIAM News* 37(4).

Barrett, C., S. Eubank, and M. Marathe. 2006. Modeling and simulation of large biological information and socio-technical systems: An interaction based approach. In *Interactive Computation: The New Paradigm*, edited by Goldin, Smolka, and Wegner. Springer-Verlag.

Barrett, C., M. Wolinsky, and M. W. Olesen. 1996. Emergent local control properties in particle hopping traffic simulations. In *Proceedings of the Conference on Traffic and Granular Flow* (TGF), edited by D.E. Wolf, M. Schreckenberg, and A. Bachem, 169–174. Singapore: World Scientific.

Beckman, R., K. A. Baggerly, and M. D. McKay. 1996. Creating synthetic base-line populations. *Transportation Research A – Policy and Practice* 30:415-429.

Bisset, K., X. Feng, M. Marathe, and S. Yardi. 2009a. Modeling interaction between individuals, social networks and public policy to support public health epidemiology. In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 2020 – 2031.

Bisset, K., and M. Marathe. 2009. A cyber-environment to support pandemic planning and response. *DOE SciDAC Magazine* 36–47.

Bisset, K., J. Chen, X. Feng, V. S. A. Kumar, and M. Marathe. 2009b. EpiFast: A fast algorithm for large-scale realistic epidemic simulations on distributed memory systems. In *Proceedings of the 23rd International Conference on Supercomputing* (ICS), 430–439. New York, NY, USA: ACM.

Bisset, K., J. Chen, X. Feng, Y. Ma, and M. Marathe. 2010. Indemics: an interactive data intensive framework for high performance epidemic simulation. In *Proceedings of the 24th International Conference on Supercomputing* (ICS), 233–242. New York, NY, USA: ACM.

Chen, J., A. Marathe, and M. Marathe. 2010. Coevolution of epidemics, social networks, and individual behavior: A case study. In Advances in Social Computing: Third International Conference on Social Computing, Behavioral Modeling, and Prediction, 218–227. Bethesda, MD, USA: Springer.

Clauset, A., C. Moore, and M. E. J. Newman. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* 453:98–101.

Contractor, N., S. Wasserman, and K. Faust. 2006. Testing multi-theoretical multilevel hypotheses about organizational networks: An analytical framework and empirical example. *Academy of Management Review* 31(3): 681.

Epstein, J. 2005. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton: Princeton University Press.

Eubank, S., H. Guclu, V. S. A. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. 2004. Modeling disease outbreaks in realistic urban social networks. *Nature* 429:180–184.

Harris, K. M. 2008. The national longitudinal study of adolescent health (Add Health), waves I and II, 1994-1996; wave III, 2001-2002 [machine-readable data file and documentation]. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

Jackson, M. O. 2007. The study of social networks in economics. In *The missing links: Formation and decay of economic networks*, edited by James E. Rauch, 19–43. New York, NY, USA: Russell Sage Foundation Publications.

Jackson, M. O., and L. Yariv. 2009. Diffusion, strategic interaction, and social structure. In *Hand- book of Social Economics*, edited by J. Benhabib, A. Bisin, and M. O. Jackson. Elsevier.

Kast, F.E., and J.E. Rosenzweig. 1981. General systems theory: Applications for organization and management. *Journal of Nursing Administration* 11(7): 32.

Leidig, J., E. Fox, M. Marathe, and H. Mortveit. 2010. Epidemiology experiment and simulation management through schema-based digital libraries. In *Proceedings of the 2nd DL.org Workshop at ECDL, Making Digital Libraries Interoperable: Challenges and Approaches*, 57–66.

Newman, M. 2003. The structure and function of complex networks. *SIAM Review* 45: 167–256.

Oreskes, N. 2000. Why believe a computer? Models, measures, and meaning in the natural world. In *The Earth Around Us: Maintaining a Livable Planet*, edited by Jill S. Schneiderman, 70–82. San Francisco: W. H. Freeman and Co.

—. 2003. The role of quantitative models in science. In *Models in Ecosystem Science*, edited by Charles D. Canham, Jonathan J. Cole, and William K. Lauenroth, 13–31. Princeton: Princeton University Press.

Oreskes, N., and K. Belitz. 2001. Philosophical issues in model assessment. In *Model Validation: Perspectives in Hydrological Science*, edited by M. C. Anderson and P. D. Bates, 23–41. London: John Wiley and Sons, Ltd.

Rice, J. A. 2006. *Mathematical Statistics and Data Analysis*. Pacific Grove, CA, USA: Duxbury Press.

Sterman, J. D. 2006. Learning from evidence in a complex world. *Am. J. Public Health* 96(3): 505–514.

Vega-Redondo, F. 2006. *Diffusion, Search and Play in Complex Social Networks*. Econometric Society Monograph Series.

Vega-Redondo, Fernando. 2007. *Complex Social Networks*. Cambridge: Cambridge University Press.

Von Bertalanffy, L. 1968. *General System Theory: Foundations, Development, Applications*. New York: G. Braziller.

Young, H. Peyton. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton: Princeton University Press.