Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

CHISIM: AN AGENT-BASED SIMULATION MODEL OF SOCIAL INTERACTIONS IN A LARGE URBAN AREA

Charles M. Macal
Nicholson T. Collier
Jonathan Ozik
Eric R. Tatara
John T. Murphy

Decision and Infrastructure Sciences
Argonne National Laboratory
9700 S. Cass Ave.
Argonne, IL 60439 USA

ABSTRACT

Cities are complex, dynamic, evolving adaptive systems comprised of people as well as interconnected physical infrastructure. Simulation modeling can help us understand and shape the evolution of our cities. In this paper, we describe an agent-based simulation modeling framework applied to Chicago, called chiSIM (for the Chicago Social Interaction Model). Each person residing in Chicago is represented as an agent in chiSIM; all places where people can be located in Chicago also are represented. The model simulates the movements of people between locations on an hourly basis during the course of a typical day. Co-located agents engage in various kinds of social interactions, such as exchanging information, engaging in business transactions, or simply sharing physical proximity. We discuss technical approaches to large-scale urban modeling including development of synthetic populations, efficiency gains through distributed processing, logging and analysis of simulation results, and visualization.

1 INTRODUCTION

“A city is first and foremost a social reactor. It works like a star, attracting people and accelerating social interactions” (Bettencourt 2013).

Cities are complex systems whose functioning depends upon many social, economic, and environmental factors (Bettencourt 2013) . Many cities, called megacities, around the world are growing exponentially in size and complexity. One hundred fifty years of rapid global urbanization have taken the global urban population from 14% (in 1990) to 50% (in 2008) and projected to be 70% by 2050 (Population Reference Bureau 2013). China alone is projected to have 221 cities with 1 million or more people by 2030, with the addition of 400 million city dwellers; between now and 2020 (Dobbs 2010), the Guangdong province will invest $229 billion in transport infrastructure to create a single 50 million person city.

There is a growing realization that cities are sources of potential opportunity as well as instability. With this realization has come the need for methods and tools that can help us understand the processes by which cities will grow and transform themselves in the coming years. Simulation has a lot to offer in understanding the cities of the future, and helping to plan for sustainable growth Many analytical and modeling techniques are finding value in developing an understanding of urban processes (Johnson et al. 2016). Computer simulation models can capture the dynamic processes operating within the city environment and offer a promising approach to addressing these and others questions. Many models have been developed for cities that focus on the lifeline physical infrastructures: transportation, energy, water, communications, and others; typically, these models ignore the interconnectivity between infrastructures and the people that

978-1-5386-6572-5/18/$31.00 ©2018 IEEE

utilize the services provided by them. An important challenge is to include people and social processes into models of cities. This paper describes a model applicable to large urban areas that includes people--their behaviors, and social interactions--as central elements in shaping a city's future.

chiSIM, the Chicago Social Interaction Model, is an agent-based model framework of people and places in Chicago along with the daily activities in which they engage. Agent-based simulation, a relatively new approach to modeling populations of heterogeneous, interacting, adaptive agents (Macal 2016) at an individual, granular level of detail, is the technical simulation approach taken in chiSIM. chiSIM models the behaviors and social interactions of all Chicago residents. Agents consist of the population of all the residents of Chicago represented at the individual level; places consist of geo-located parcels in the city, such as households, schools, workplaces, hospitals, and general quarters, such as nursing homes, dormitories, jails, etc. During the course of a simulated day, agents move from place-to-place, hour-by-hour, engaging in social activities and interactions with co-located agents.

The chiSIM framework has many potential uses such as forecasting the need for transportation, energy and other infrastructure services; the spread of infectious diseases; the spread of information; the adoption of new technologies; the effectiveness of social programs; and many other uses. To date, chiSIM has been used to understand what measures could be used to mitigate the spread of community-associated MRSA (Methicillin-resistant *Staphylococcus aureus*), a skin and soft tissue infection caused by *Staphylococcus aureus* bacteria (Macal et al. 2014), and a community healthcare information program (Kaligotla et al. 2018). Work in progress includes modeling the transportation needs of the population on the infrastructure including modeling charging station location (electric vehicles) and fueling station location (hydrogen-powered vehicles), HIV, Hepatitis C and other infectious disease transmission in high-risk communities, and forecasting the individual decisions on residential energy consumption and the consumer adoption of residential rooftop solar energy technologies.

2 BACKGROUND

Modeling cities and their populations has a long history, originating with the invention and rising popularity of simulation as a field, chronicled by Batty (2008), who also notes that urban modeling was driven in part by the availability and increasing power of computers, though these early efforts often exceeded the computing power available. Early urban models, in the 1950's and 1960's onward, were systems dynamics approaches (Forrester 1969). Cellular automata (CA) then enjoyed a heyday in the late 1980's and 1990's and eclipsed SD models at least in terms of the extent of the literature (Batty 1989). Most of the modeling effort uses CA to examine urban growth or land use change, and the systems dynamics approach are broadly economic (industry output, labor pools, etc.).

Other approaches like 3-D modeling and addressing urban research through "big data" approaches are becoming more prominent. Cities have been studied from the standpoint of complexity. One research program focuses on general or universal principles characterizing cities and their macroscopic behaviors. For example, scaling laws have been derived to predict the average social, spatial, and infrastructural properties of cities as a set of scaling relations. There are also system- or infrastructure specific urban simulations focusing fir example on traffic, pedestrian movement, disease spread, disaster response and evacuation, etc.

Agent-based models (ABM) were introduced around the early 2000's. ABMs have been applied to a very broad set of applications including: traffic and transportation, pedestrian movement, social contact networks, crime, segregation, disease, natural resource management and sustainability, economics and real estate, disasters, and others. We use an agent- based modeling approach in chiSIM because the agent approach can best represent the diversity of an entire population in terms of the characteristics and behaviors of individuals. Few if any urban models exist at the large scale of the chiSIM, in terms of the number of individuals and behaviors considered at the fine-grained level of geo-spatial detail for such a large geographical area.

3 SIMULATING PEOPLE, PLACES, AND ACTIVITIES

Assembling the large amount of information from a variety of heterogeneous data sources to support a city-scale ABM is a challenging technical and logistical problem. An agent-based city model begins with a synthetic population of individuals (people) who are represented as agents in the model (Figure 1). There is no data on actual, specific individuals in the model but characterizations of individuals that are accurate when aggregated. Synthetic populations with baseline socio-demographic data are derived from combined U.S. Census files that are available from a growing number of sources. chiSIM uses baseline synthetic population data such as those developed by Wheaton et al. (2009). The socio-demographic attributes of the synthetic population match that of the actual population for Chicago in the aggregate for the Census years of 2000 and 2010. Each agent has a baseline set of socio-demographic characteristics (e.g., race/ethnicity, age, gender, educational attainment, income).

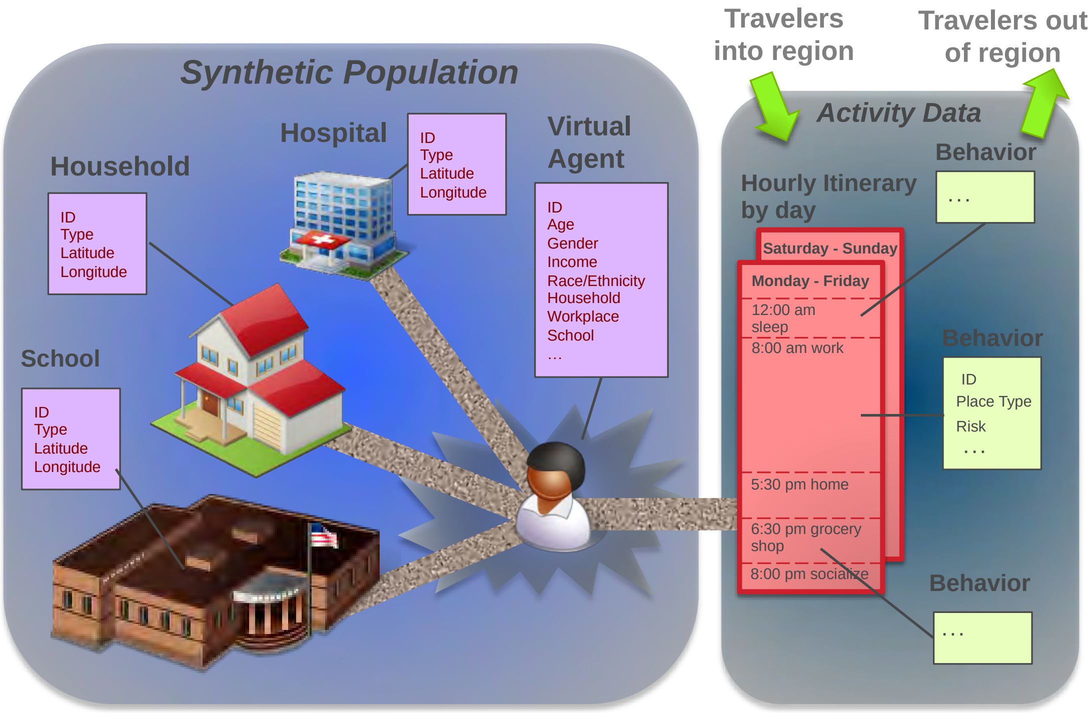

Figure 1: chiSIM agents.

An agent-based city model also includes a synthetic population of places (Figure 2). All places are characterized by place type, including households, schools, hospitals and workplaces, and have locations. The synthetic population assigns agents to households, workplaces and schools (for those of school age). Places are categorized as having different types of activities that may occur there.

We combine several publicly available national data sources to model activities of the synthetic agent population. Each agent has a daily activity profile that determines what times throughout the day he or she occupies each location. Activity profiles are empirically based on 24-hour time diaries collected as part of the U.S. Bureau of Labor Statistics' annual American Time Use Survey (ATUS) for individuals aged 15 years and older and from the Panel Study of Income Dynamics (PSID) for children younger than 15 years. Both are nationally representative samples and collect diary data on randomly assigned days. The diary records each activity during the 24-hour period (start/stop times, location and others present). Two profiles (one weekday and one weekend) from respondents living in metropolitan areas are assigned to each agent in the model. This is done by stochastically matching each agent with an ATUS or PSID respondent who is either identical or similar with respect to socio-demographic characteristics.

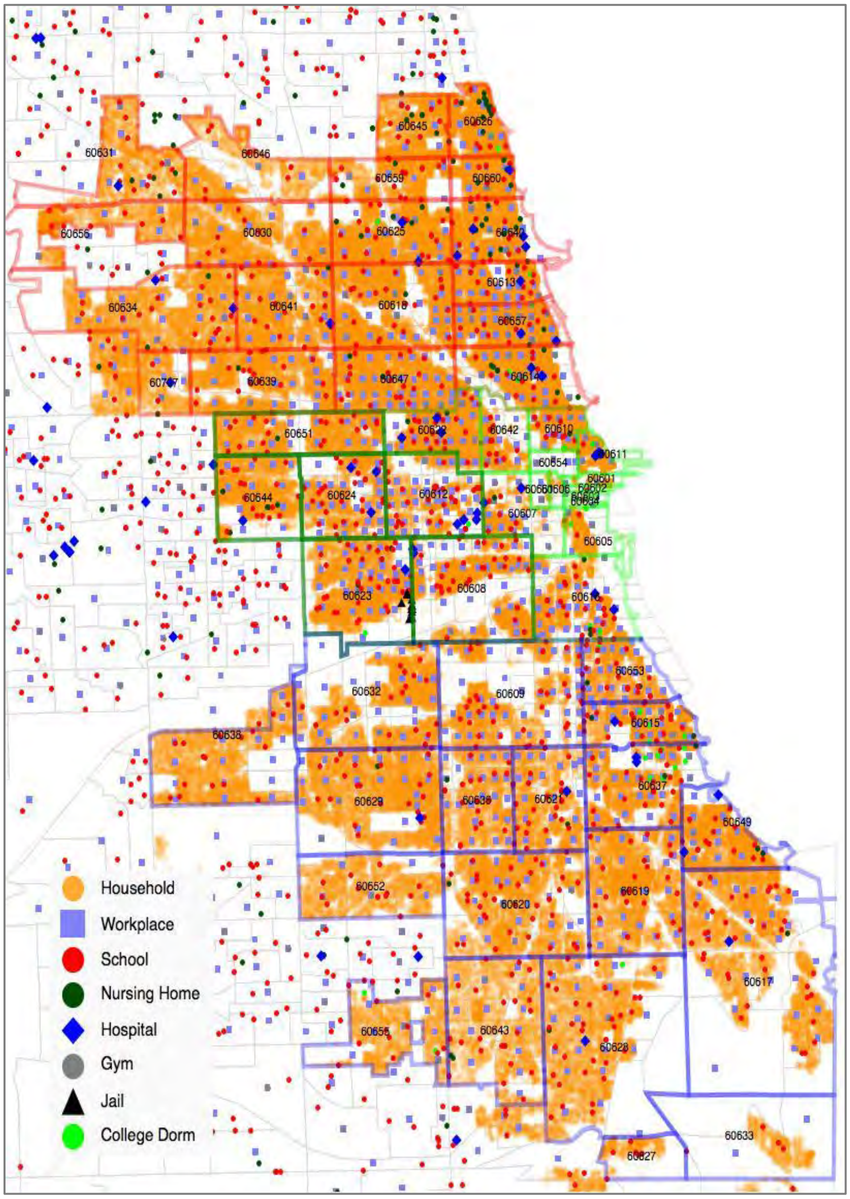

Figure 2: chiSIM map of places in Chicago.

Basic agent behaviors, which can be gleaned from activity data, affect activity choice (what activities do agents engage in at any time), and movement (when and where do agents change their location to engage in a new activity). More specialized behaviors are application specific, including communication (what specific information is exchanged in social interactions and with whom) and actions in unusual or specific circumstances. For example, in the community health information study, a sampling of healthcare recipients were surveyed on who and how often they pass on information from healthcare providers regarding healthcare services available in their community. In a pandemics simulation, a behavioral ontology was developed from the literature that enumerated actions people take in the event of a pandemic and the circumstances under which those actions were taken. The behavioral ontology then was implemented in the agent-based model to drive agent behaviors.

4 CHISIM FRAMEWORK OVERVIEW

The chiSIM frameworks enables building hybrid time-stepped and discrete-event, stochastic simulations. The prototypical chiSIM model moves 3 million individual agents to and from 1.2 million places in Chicago on an hourly basis. The simulation is initialized at hour zero with agents located across Chicago. The Generalized Mobility/Activity Model advances agent locations appropriate for the hour of the day, moving agents between their own households, schools, workplaces, other households, etc. Specialized mobility/activity models are developed separately to model the circulation of agents from the community into and out of special places that are not included in the general activity profile data, since that is developed for general populations. Once activity state changes are recorded, or logged, the hour is incremented and the simulation processes are repeated until the simulation reaches the end time, at which point yearly and hourly summary reports are produced. chiSIM has simulated agent movement and activities by hour over 10 years. For example, to explore possible explanations for the ongoing MRSA epidemic in Chicago, the model was simulated for the period 2001-2010 and was able to reproduce the empirical buildup that was observed. For the 10-year hurly simulation (It was estimated that over 2 trillion total people-to-people contacts were simulated over the 87,600 hourly time steps in the simulation.)

A large-scale agent-based model such as chiSIM presents challenges for computational resources. For example, the MRSA ABM originally ran for 60 hours on a single compute node. Long run times for simulations may severely limit the usefulness of a model for policy analysis. Collier et al. (2015) explored the benefits of distributing a model across multiple computing nodes. They ran a series of experiments that exploited the location/movement network structure by partitioning locations into separate computing processes based on the expected people flow between the locations. Compute node thread parallelization using OpenMP and distributed parallelization across multiple processes using MPI reduced the run time to 4 hours utilizing 128 compute nodes. Other distributed computing strategies promise to yield additional benefits.

One of the challenges of working with a complex agent-based model is to record the “history” that the model creates. The history records all the agent events by location and time over the course of the simulation. Information on every agent’s state and location, and all the other agents that an agent is having contact with at every moment, is known in the simulation. The simulation can be critical source of information that is otherwise unobserved or even unobservable. For example, in the MRSA model, two disease states are critical for transmission of the disease: the infected state, when an open skin or soft tissue infection is readily observable, and the colonized stated, which is asymptomatic. A colonized individual can transmit the disease but does not know they are contagious and can pass on the disease. The simulation produced a reasonable recreation of the profile of the number of infected individuals (observable, and based on empirical data) with MRSA; and the model was able to infer the correlated number of individuals that were also colonized, which was not observed and for which there no empirical data exists.

The amount of data produced by the agent-based model of millions of interacting individuals moving among millions of locations can be enormous. For example, Figure 3 shows the occupancy patterns of agents by location at specific times. Figure 4 shows movement patterns for selected agents between consecutive hours of the day. Figure 5 shows simulated disease transmission as a result of personal contact between individuals. Because the log files can be enormous in size, collecting all the agent event data over the simulation period, special techniques are required to record and analyze such large-scale datasets (Tatara et al. 2017). Figure 6 shows the results of such an analysis, a co-location network for all agents in the model who have two degrees of separation. This information is useful in infectious disease applications for tracing agent contact patterns back to the original source, effectively identifying “patient zero.”

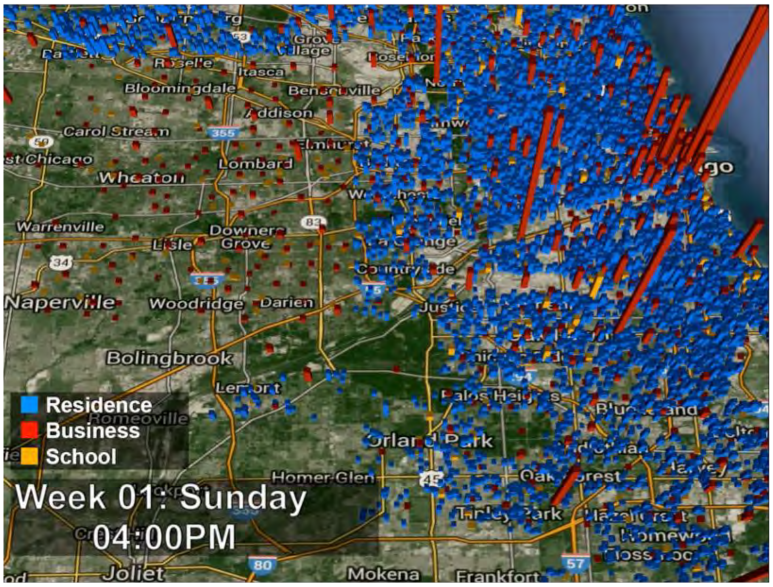

Figure 3: City occupancy by agents.

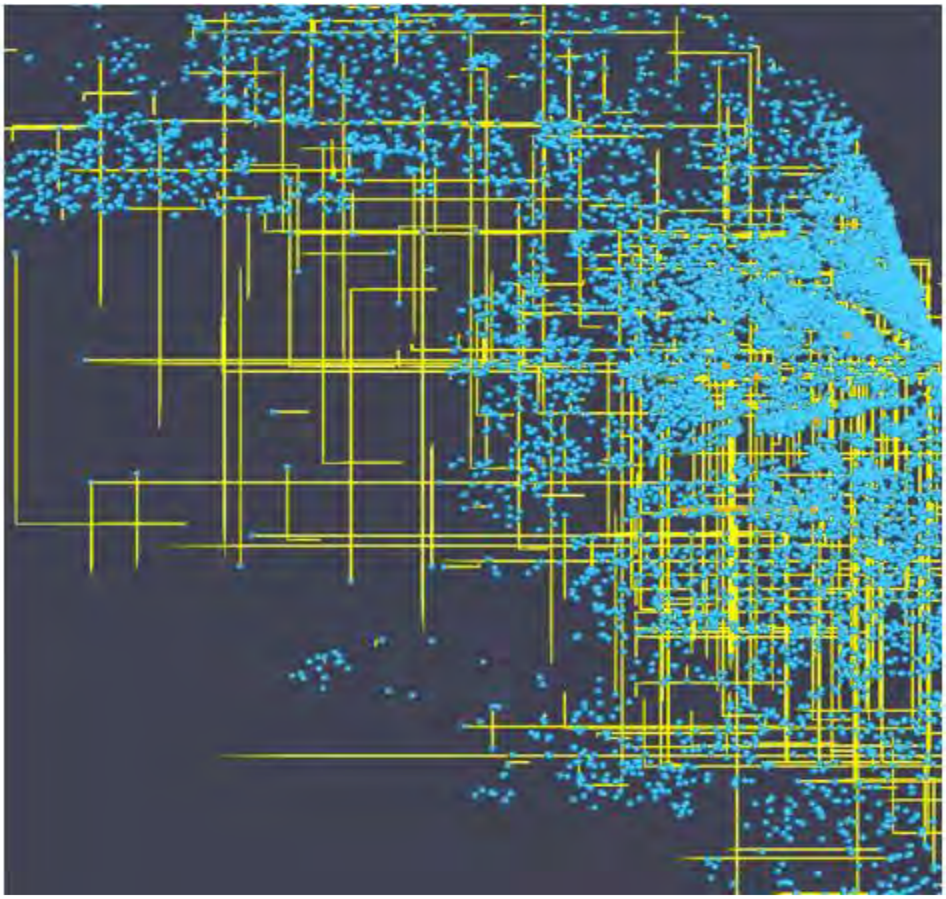

Figure 4: Agent movement patterns.

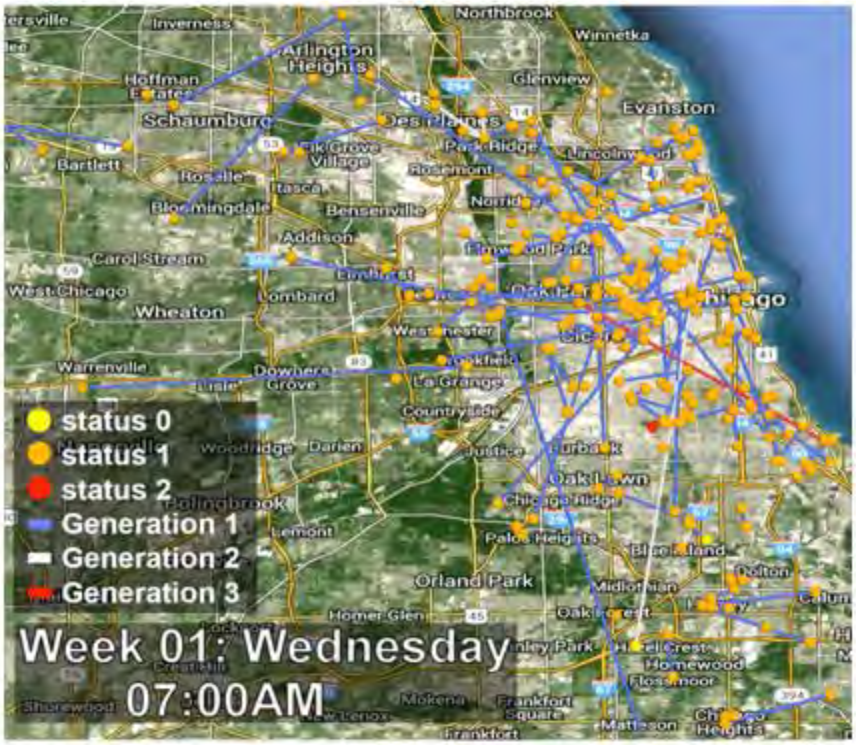

Figure 5: Agent disease transmission tracing in chiSIM.

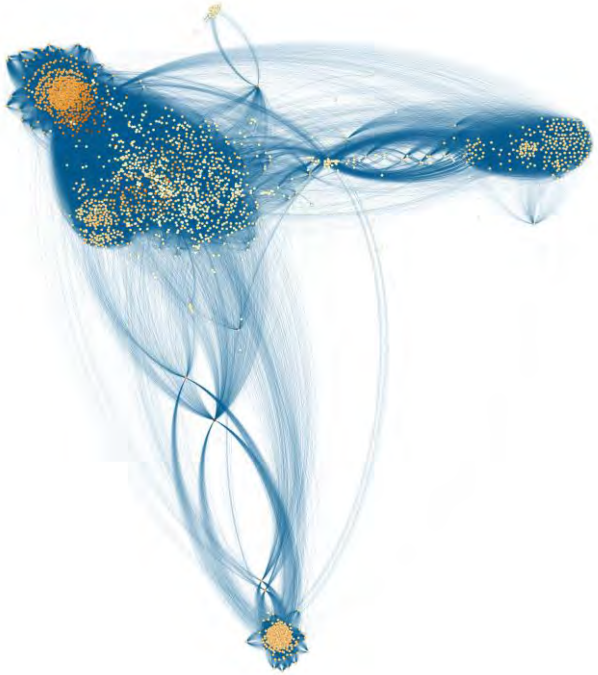

Figure 6: Sample co-location network for randomly sampled individuals and all individuals within two degrees of separation. Lines represent collocation between person nodes. Node color indicates vertex degree, with darker nodes having a greater degree than lighter nodes.

5 TECHNICAL CHALLENGES IN SIMULATING A CITY

One of the primary challenges in implementing a large city-size simulation is achieving sufficient computational performance such that model run time and memory use does not limit the utility of the model itself. Implemented in C++, using the Repast for High Performance Computing (Repast HPC) (Collier and North 2013), an agent-based model framework for implementing distributed agent-based models on high performance distributed-memory computing platforms, chiSIM achieves such performance by distributing a simulation across multiple processes.

As noted above, chiSIM began as a generalization of the model of community associated methicillin-resistant Staphylococcus aureus (CA-MRSA) (Macal et al. 2014), mentioned above in section 3. The CA-MRSA model was a non-distributed model in which all the model components, including all the people and places required to simulate the city of Chicago, ran on a single computational process and thus often pushed both memory bounds of individual computing nodes and run time limits of shared computing resources. chiSIM retains and generalizes the social interaction dynamics of the CA-MRSA model and allows models implemented using chiSIM to be distributed across multiple processes. Places are created on a single process and remain there. Persons move between the processes according to their activity profiles, which are generated from a problem-specific activity model describing how persons move about the community.

When a person selects a next place to move to, the person may stay on its current process or it may have to move to another process if its next place is not on the current process. This cross-process distribution ameliorates both run time and memory issues. By having fewer people and places on each process, and allowing for processes to run in parallel, the time to iterate through each person's and place's behavior is reduced. *N* number of people can be split in to *N / number of processes* sets of people, each of which runs in parallel. Similarly, fewer people and places on a process also reduces an individual process' memory requirements.

However, parallelizing and distributing code across multiple processes is not without its own performance pitfalls. Transferring data, that is, the internal state of people in chiSIM, between processes can be computationally expensive. In a previous paper, we describe how we used the Metis graph partitioning application to assign places to processes to minimize this computationally expensive cross-process movement of persons and to balance the number of persons on each process (Collier et al. 2015). In addition, chiSIM also provides the ability to cache any constant person state, given sufficient memory, lessening the amount of data transferred between processes. When a person is moved between processes only the dynamic state needs to be moved, any constant state can be retrieved from the cache when the person is unpacked by the retrieving process.

As a software framework, chiSIM provides both the structure and an application programming interface (API) for implementing, and more significantly distributing, a city-sized simulation across multiple processes. The API allows the user to easily implement cross-process person movement as well as the code that drives such movement, such as via activity schedules. In addition, it also provides flexible scheduling and data collection (via Repast HPC), utility code for reading input files, working with model parameters and other simulation implementation tasks, as well as providing some example simulations. A user of chiSIM needs to provide the synthetic population, places and activity schedules, and implement any model-specific behavior for persons, places, and activities, using chiSIM provided base classes and utility code.

The software stack for a typical chiSIM-based model is shown in Figure 7. For example, given a city-sized epidemiological simulation in which people spread a disease through co-location with other people, the user needs to provide as input data the synthetic population, the places in which people can be co-located, and activity schedules that describe when the people move to those places. The model specific behavior that needs to be implemented are such things as disease state transitions within each person, the disease transmission between co-located people, and the selection of a destination place given an activity schedule. The latter can be as simple as a direct mapping between time and place (e.g., at nine o'clock go to work) or a more complex dynamic decision that may depend on additional person and simulation state.

These model-specific components are built on chiSIM provided code, and integrated into the chiSIM provided structure to create the working model. The chiSIM framework is free and open source, and is available at github.com/Repast/chiSIM. The chiSIM repository contains detailed instructions for building and using the framework, along with example chiSIM-based simulations.

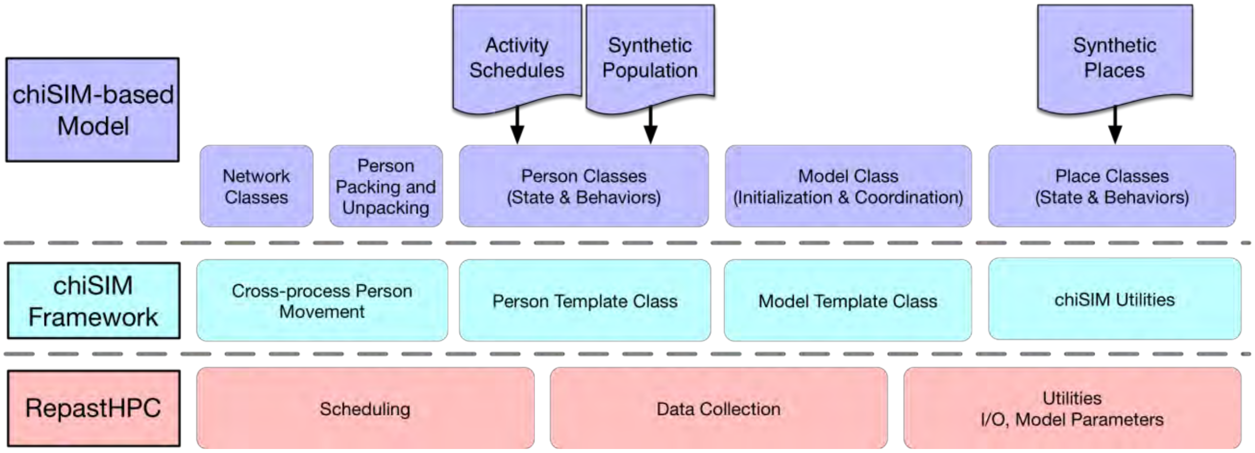
chiSIM-based Model
Activity Schedules
Synthetic Population
Synthetic Places
Network Classes
Person Packing and Unpacking
Person Classes (State & Behaviors)
Model Class (Initialization & Coordination)
Place Classes (State & Behaviors)
chiSIM Framework
Cross-process Person Movement
Person Template Class
Model Template Class
chiSIM Utilities
RepastHPC
Scheduling
Data Collection
Utilities I/O, Model Parameters

Figure 7: The software stack for a chiSIM-based model.

A chiSIM model enables *in silico* experimentation of city-scale phenomena. However, due to the stochastic processes within a typical model and its likely large parameter space, tasks such as model calibration and validation, characterization of the model input parameter space, propagating of input data uncertainties, and the computational discovery of effective interventions, tasks needed for both creating a trusted model and utilizing it, require the ability to run large ensembles of model instances based on complex and iterative logic, what we term *model exploration* (ME). Examples of ME algorithms include Approximate Bayesian Computing (Beaumont 2010), Active Learning (Settles 2012), and genetic/evolutionary algorithms (Holland 1992). Given the potentially significant computation involved in running a single distributed chiSIM model, coordinating an algorithm that dynamically generates *collections* of models based on prior iterations is a non-trivial endeavor. To address the need for running large-scale ME experiments, we have developed the Extreme-scale Model Exploration with Swift (EMEWS) framework (Ozik et al. 2016) to enable the creation of automated, iterative HPC workflows. EMEWS, built on top of the Swift/T parallel scripting language (Wozniak et al. 2013), enables the framework user to directly plug in a distributed MPI-based model and an existing ME algorithm implementing potentially complex iterating logic to create a large-scale HPC workflow. ME algorithms can be expressed in popular data analytics languages such as Python and R. Despite its flexibility an EMEWS workflow is highly performant and scalable to the largest HPC resources. Large EMEWS workflows of chiSIM models have been run on a variety of computing resources, including the Mira BlueGene/Q supercomputer, Cray XE6 Beagle2 supercomputer, Blues and Bebop clusters at Argonne National Laboratory and the Midway2 cluster at the University of Chicago.

6 SUMMARY AND CONCLUSIONS

Given the expected continuing rapid increases in urban development, the need, as well as the opportunities, for tools to understand and proactively plan the orderly development of sustainable cities is greater than ever. Recent and ongoing developments in computing capabilities, both in terms of software and hardware, the advancement of new modeling paradigms, such as agent-based modeling, combined with the ever-increasing availability of big data provide unprecedented opportunities for modeling and simulation to make important and broad impacts on cities. Models can contribute in backcasting (explaining the past), forecasting the future, testing proposed interventions, or explicitly optimizing urban processes. With models as experimental vehicles and model exploration frameworks facilitating the use of high-performance computing resources many experiments can be run rapidly and much can be learned quickly. Still, many challenges remain, as this paper has outlined, from assembling large datasets from disparate sources, to computational efficiency challenges, to logging models and handling the analytical challenges that large datasets produced by simulation models generate.

ACKNOWLEDGEMENTS

This work is supported by the U.S. Department of Energy under contract number DE-AC02-06CH11357 and NSF RAPID Award DEB-1516428. This research used computing resources provided on Blues, a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory, which is a DOE Office of Science User Facility.

REFERENCES

Batty, M., P. Longley, and S. Fotheringham. 1989. "Urban Growth and Form: Scaling, Fractal Geometry, and Diffusion-Limited Aggregation". *Environment and Planning* 21(11):1447–1472.

Batty, M. 2008. "Fifty Years of Urban Modeling: Macro-Statics to Micro-Dynamics." Edited by Sergio Albeverio, Denise Andrey, Paolo Giordano, and Dr. Alberto Vancheri. *The Dynamics of Complex Urban Systems: An Interdisciplinary Approach*. Physica-Verlag HD. http://link.springer.com/chapter/10.1007/978-3-7908-1937-3_1.

Bettencourt, L. M. A. 2013. "The Origins of Scaling in Cities". *Science*, 340(6139):1438-1441.

Beaumont, M. A. 2010. "Approximate Bayesian Computation in Evolution and Ecology". *Annual Review of Ecology, Evolution, and Systematics* 41(1):379-406.

Collier, N. T., J. Ozik, and C. M. Macal. 2015. "Large-scale ABM with RepastHPC: A Case-study in Parallelizing a Distributed ABM". in *PADABS 2015 Proceedings*. Vienna, Austria, 24-28 August 2015.

Collier, N., and M. North. 2013. "Parallel Agent-based Simulation with Repast for High Performance Computing". *Simulation* 89 (10): 1215–1235.

Dobbs, R./McKinsey Global Institute. 2010. "Megacities". *Foreign Policy Magazine* .

Forrester, Jay W. 1969. *Urban Dynamics*. 1st ed. Pegasus Communication Inc, Cambridge, MA.

Holland, J. H. 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. 1st ed. Cambridge, Mass: Bradford.

Johnson, R., P. Sydelko, C. Macal, J. Hummel, and M. North. 2016. "Geospatial Analytics Applied to Urban Environments, *Workshop: Mad Scientist 2016: Megacities and Dense Urban Areas in 2025 and Beyond*, April 21-22, 2016, Tempe, Arizona.

Kaligotla C., J. Ozik, N. Collier, C. M. Macal, S. Lindau, E. Abramsohn, and E. Huang. 2018. "Modeling an Information-Based Community Health Intervention on the South Side of Chicago". *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe et al., Piscataway, New Jersey:IEEE.

Macal, C. M. 2016. "Everything You Need to Know About Agent-based Modeling and Simulation". *Journal of Simulation* 10:144. doi:10.1057/jos.2016.7.

Macal, C. M., M. J. North, N. Collier, V. M. Dukic, D. T. Wegener, M. Z. David, R. S. Daum, P. Schumm, J. A. Evans, J. R. Wilder, L. G. Miller, S. J. Eells, and D. S. Lauderdale. 2014. "Modeling the transmission of community-associated methicillin-resistant *Staphylococcus aureus*: a dynamic agent-based simulation". *Journal of Translational Medicine* 12:124.

Ozik, J., N. T. Collier, J. M.. Wozniak, and C. Spagnuolo. 2016. "From desktop to large-scale model exploration with Swift/T". In *Proceedings of the 2016 Winter Simulation Conference*, 206-220 , edited by S.Jain et al., Piscataway, New Jersey:IEEE.

Population Reference Bureau. 2013. *Food Insight*.

Settles, B. 2012. “Active Learning”. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

Tatara, E., N. Collier, J. Ozik, and C. Macal. 2017. “Endogenous Social Networks from Large-Scale Agent-Based Models”. *IEEE Workshop on Parallel and Distributed Processing for Computational Social Systems (ParSocial 2017),* June 2, 2017, Orlando, FL USA.

Wheaton, W. D., et al. 2009. *Synthesized Population Databases: A US Geospatial Database for Agent-Based Models* RTI Press publication No. MR-0010-0905.

Wozniak, J. M., T. G. Armstrong, M. Wilde, D. S. Katz, E.. Lusk, and I. T. Foster. 2013. “Swift/T: Large-Scale Application Composition via Distributed-Memory Dataflow Processing.” In *Cluster, Cloud and Grid Computing (CCGrid)*, 2013 13th IEEE/ACM International Symposium, 95–102. IEEE.

AUTHOR BIOGRAPHIES

CHARLES M. MACAL, PhD, PE, is the Director of the Social, Behavioral and Decision Sciences group at Argonne National Laboratory. He is a member of the INFORMS-Simulation Society, Association for Computing Machinery, the Society for Computer Simulation International, and the System Dynamics Society. He has a Ph.D. in Industrial Engineering & Management Sciences from Northwestern and a Master’s Degree in Industrial Engineering from Purdue. Contact: macal@anl.gov.

NICHOLSON COLLIER PhD, is Software Engineer in the Decision and Infrastructure Sciences Division at Argonne National Laboratory and Research Staff at the Computation Institute at the University of Chicago. He is the lead developer of the Repast and RepastHPC(https://repast.github.io) agent-based modeling toolkits and a core developer of the EMEWS (http://emews.org) framework for large-scale model exploration. Contact: nick.collier@gmail.com.

JONATHAN OZIK PhD, is a Computational Scientist at Argonne National Laboratory and a Senior Fellow at the Computation Institute at the University of Chicago. He leads the Repast agent-based modeling toolkit (repast.github.io) and the EMEWS framework (emews.org) projects. His research focuses on agent-based modeling and large-scale model exploration. Contact: jozik@anl.gov.

ERIC R. TATARA, PhD, is a Software Engineer at Argonne National Laboratory. He has degrees in chemical engineering. Contact: tatara@anl.gov.

JOHN T. MURPHY, PhD, is a Computational Social Scientist at Argonne National Laboratory. He has degrees in anthropology. Contact: jtmurphy@anl.gov.