The Journal of Infectious Diseases

SUPPLEMENT ARTICLE

epiDMS: Data Management and Analytics for Decision-Making From Epidemic Spread Simulation Ensembles

Sicong Liu,[1] Silvestro Poccia,[3] K. Selçuk Candan,[1] Gerardo Chowell,[2] and Maria Luisa Sapino[3]

[1]School of Informatics, and Decision Systems Engineering, Arizona State University, Tempe; [2]School of Public Health, Georgia State University, Atlanta; and [3]Computer Science Department, University of Torino, Italy

Background. Carefully calibrated large-scale computational models of epidemic spread represent a powerful tool to support the decision-making process during epidemic emergencies. Epidemic models are being increasingly used for generating forecasts of the spatial-temporal progression of epidemics at different spatial scales and for assessing the likely impact of different intervention strategies. However, the management and analysis of simulation ensembles stemming from large-scale computational models pose challenges, particularly when dealing with multiple interdependent parameters, spanning multiple layers and geospatial frames, affected by complex dynamic processes operating at different resolutions.

Methods. We describe and illustrate with examples a novel epidemic simulation data management system, epiDMS, that was developed to address the challenges that arise from the need to generate, search, visualize, and analyze, in a scalable manner, large volumes of epidemic simulation ensembles and observations during the progression of an epidemic.

Results and conclusions. epiDMS is a publicly available system that facilitates management and analysis of large epidemic simulation ensembles. epiDMS aims to fill an important hole in decision-making during healthcare emergencies by enabling critical services with significant economic and health impact.

Keywords. epidemics; big data; simulation ensembles; data management; analytics; public health decision-making.

The potential for pandemics to rapidly generate morbidity and mortality and influence economies around the world has highlighted the need to develop quantitative frameworks for supporting public health decision-making in near real time. For instance, the 2003 SARS coronavirus emergency, which originated in China and spread to 29 countries, generated important nosocomial outbreaks in several regions by August 2003 [1, 2]. More recently, the 2009 influenza A(H1N1) pandemic, originating in Mexico, rapidly spread around the globe via the airline network and reached 20 countries, with the highest volume of passengers arriving from Mexico within a few weeks of epidemic onset [3]. Importantly, the economic cost associated with a pandemic similar to the 2009 influenza A(H1N1) pandemic has been estimated to range from $360 billion to $4 trillion [4] for the first year of virus circulation.

Large-scale computational transmission models of infectious disease spread are increasingly becoming part of the toolkit to generate inferences about the spread and control of infectious diseases. Examples of real-time analyses of epidemics supported by large-scale transmission models include estimating the transmissibility of an epidemic disease, such as influenza [5–7]; forecasting the spatiotemporal evolution of pandemics at different spatial scales [8, 9]; assessing the effect of travel controls during the early epidemic phase [10–12]; predicting the effect of school closures in mitigating disease spread [13–15]; and assessing the impact of reactive vaccination strategies [16]. These analyses, however, require access to, integration of, and analysis of models and large volumes of data, including data sets from diverse sources, to parameterize demographic characteristics, contact networks, age-specific contact rates, mobility networks, and healthcare and control interventions.

In this article, we argue that, if effectively leveraged, existing simulation analyses and real-time observations generated during an outbreak can be collectively used for better understanding the transmission dynamics and refining existing models. At the same time, these model simulations are useful for performing exploratory, if-then types of hypothetical analyses of epidemic scenarios to address critical questions, including whether we can identify and classify key events (eg, epidemic peak timing and likely epidemic duration) during an infectious disease outbreak from large simulation ensembles, compare and summarize a large number of epidemic simulations and observations under different epidemiological scenarios, and discover latent relationships and dependencies among disease dynamics and social parameters.

EPIDEMIC SIMULATIONS

Global epidemic spread can be characterized via simulation through networks of multiple (local and global) scales: individuals within a subpopulation may be infected through local contacts during a localized outbreak. These infected individuals

Correspondence: K. S. Candan, Computer Science and Engineering, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287-8809 (candan@asu.edu).

The Journal of Infectious Diseases® 2016;214(S4):S427–32

DOI: 10.1093/infdis/jiw305

then may seed the infection in other regions, starting a new outbreak. Thus, large-scale epidemic simulation systems (eg, GLEaM [9] and STEM [17]) are required to leverage models and data at different spatial scales. These include social contact networks, local and global individual mobility patterns, location-specific control interventions, and epidemiological characteristics of the infectious disease in question.

The population model for a global epidemic simulation system can be based, for example, on the Gridded Population of the World project by the Socioeconomic Data and Applications Center [18], which has a resolution of 15×15 minutes of arc.

Mobility models can include long-range air travel mobility data, such as those from the International Air Transport Association and the Official Airline Guide, and/or short-range commuting patterns between adjacent subpopulations. High-resolution demographic and age-specific contact data have become available for a number of areas, including the United States [19] and Southeast Asia [16], while age-specific contact rates have been derived from population surveys for a number of European countries [20]. Large-scale computational transmission models, parameterized with high-volume air traffic data and country-level seasonality factors, are being increasingly used to assess the global transmission patterns of emerging infectious diseases and the effectiveness of control measures [21–23].

Epidemic models allow the user to specify epidemiological parameters that are specific to the infectious disease (such as transmissibility and seasonality), initial outbreak conditions (eg, the seeding characteristics of the epidemic and the immunity profile of the subpopulation), and the timing, type, and intensity of intervention measures. While the disease model can be specific to the type of infectious disease, the parameters of a typical model (eg, the modified susceptible-latent-infectious-recovered model described in ref. [9]) include (1) the infection rate of contracting illness when an individual interacts with an infectious person, (2) infection rate scaling factors for asymptomatic infectors and treated infectors, (3) the average length of the latency period (in which the individual is infected but not infecting), (4) the probability of symptomatic versus asymptomatic infections, (5) the change in travelling behavior after the onset of symptoms, (6) the average length of recovery, (7) the percentage of infectious individuals who undergo pharmaceutical treatment, and (8) the impact (eg, on the length of the infectious period) of the treatment.

The output of a simulation is a multivariate time series, which tracks for each spatial location (such as the US states) the simulation values of each output parameter, such as the number of infected individuals.

CHALLENGES

While large-scale epidemic simulation systems such as GLEaM [9] or STEM [17] represent very powerful and highly modular and flexible epidemic spread simulation systems, their power for real-time decision-making could be enhanced by addressing two challenges. First, a sufficiently useful disease spreading simulation system requires complex models, including social contact networks, local and global mobility patterns of individuals, and epidemiological parameters for the infectious disease (eg, the infectious period). Epidemic simulations track tens or hundreds of interdependent parameters, spanning multiple layers and geospatial frames, affected by complex dynamic processes operating at different resolutions. Moreover, an ensemble of stochastic epidemic realizations may include hundreds or thousands of simulations, each with different parameters settings corresponding to slightly different but plausible scenarios [24, 25]. As a consequence, running and interpreting simulation results (along with the real-world observations) to generate timely actionable results pose challenges.

A second a major challenge in using data- and model-driven computer simulations for predicting geotemporal evolution of epidemics for managing health emergencies, such as the 2014–2015 Ebola epidemic in West Africa, is that the data, models, and underlying model parameters dynamically evolve over time. This necessitates continuous analyses and interpretations of the incoming data and adaptation of the networks and models. Therefore, simulation ensembles may need to be continuously revised and refined as the situation on the ground changes. Revisions involve incorporating the real-world observations, as well as updated probability surfaces, into existing simulations to alter their outcomes, whereas refinements involve identifying new simulations to run based on the changing situation on the ground to provide trustable recommendations. As the situation on the ground and intervention mechanisms evolve, the sampling strategies for the input parameter spaces have to be varied (by eliminating irrelevant scenarios and considering new scenarios or varying the likelihood of old scenarios) in such a way that more-accurate simulation results are obtained where it is more relevant.

To have a significant impact on disease control and to devise validated epidemic response strategies within a realistic time frame, public health authorities need to adequately and systematically interpret observations, understand the processes driving epidemic outbreaks, and assess the robustness of conclusions driven from simulations. Because of the volume and complexity of the data, the varying spatial and temporal scales at which the key transmission processes operate and relevant observations are made, public health experts could benefit from novel decision support systems. Therefore, tools that help execute large-scale simulation ensembles under a large number of diverse hypotheses/scenarios and those that facilitate analysis, exploration, interpretation, and visualization of large simulation ensembles (aligned with the real-world observations) to generate timely actionable results are critically needed for understanding the evolution patterns of the outbreaks (including

estimating transmissibility, forecasting the spatiotemporal spread at different spatial scales, and assessing the cost and impact of interventions, including travel controls, at various stages of the epidemic) and supporting real-time decision-making and hypothesis testing through large-scale simulations.

epiDMS OVERVIEW AND USE SCENARIO

Data and models relevant to data-intensive simulations are voluminous, multivariate, have multiple resolutions, multilayered, geotemporal, interconnected and interdependent, and often incomplete/imprecise. Moreover, data and models dynamically evolve over time, owing to control actions taken by individuals and public health interventions, requiring continuous adaptation and repeat modeling.

epiDMS, a novel epidemic simulation data management system software framework [26], aims to address the key challenges underlying large epidemic spread simulations, which, today, hinder real-time and continuous analysis and decision-making during ongoing outbreaks. Unlike other dynamic modeling platforms, such as Berkeley Madonna [27], the services provided by epiDMS include (a) storage and indexing of large-ensemble simulation data sets and the corresponding models and (b) search and analysis of ensemble simulation data sets to enable ensemble-based decision support [28–30].

The target user group for epiDMS includes a range of public health researchers and decision-makers. While creation of models for ensemble simulations and query formulation require moderate infectious disease modeling experience, epiDMS also provides parameterized queries and other interactive user interfaces to enable decision-makers with minimal experience to explore large-ensemble simulations.

System Overview

epiDMS [26] consists of three major components for managing the data and models for data-driven real-time epidemic simulations (Figure 1). First, the epidemic ensemble execution engine (epiRun) takes as input an epidemic model, mobility/connectivity models, interventions, and outbreak conditions (such as ground zero) and creates an epidemic ensemble by sampling the disease parameter space and executing simulations, using an external simulation engine. Note that epiRun is not specific to any disease model or simulation engine and that it can wrap, as a black-box software component, any epidemic simulation engine as long as it provides command line invocation. The epidemic model (formulated in the format specific to the simulation engine), the selected input parameter values, and the simulation results (ie, the time series for each output variable) then become inputs for the epidemic data and model store (epiStore).

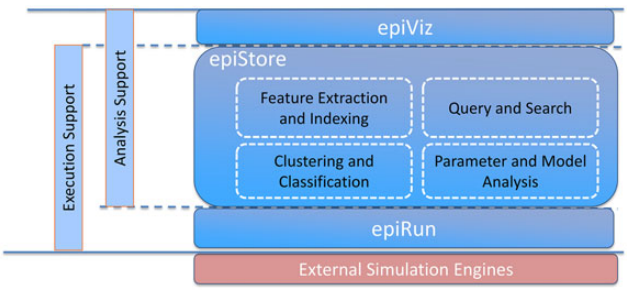

Figure 1. epiDMS overview.

Second, epiStore stores and indexes the relevant data and metadata sets. The data and models relevant for modeling large-scale epidemics include the following: one or more network layers for epidemic simulation, from local and global mobility patterns to social contact networks; disease models, which describe the epidemiological parameters relevant to a simulation and the parameter dependencies necessary in the computation of the disease spread; time series, from different simulations, each corresponding to different sets of assumptions (disease parameters or models) or context (eg, spatiotemporal context, outbreak conditions, or interventions); and disease observations, which include real-world observations that arise in near real time relating to a particular epidemic, including the spread and severity of the disease and observations about other relevant parameters, such as the average length of recovery or percentage of infectious individuals that undergo pharmaceutical treatment. epiStore captures simulation metadata (ie, simulation model, parameter values, and connectivity graphs) and simulation outputs (ie, time series) and provides data analysis (eg, clustering, classification, and event extraction) to support decision-making. Once again, epiStore is not specific to any disease model or simulation ensembles generated by a specific simulation engine—it can read and store models and simulation results produced by any epidemic simulation engine as long as data wrappers that convert data and metadata into internal epiStore representation are available.

Third, the epidemic ensemble query, visualization, and exploration module (epiViz) provides a web-based query and result visualization interface to support user interaction and exploratory decision-making through simulation ensembles (Figure 2). Query specification language is also model independent, in the sense that the system does not make any assumptions regarding what the input and output parameters of the simulations are—once imported into epiStore, parameters of any model can be queried, visualized, and explored.

epiDMS Use Scenario

Consider a governmental agency charged with developing a preparedness plan for the next influenza pandemic. To account for uncertainty in the epidemiology of the disease, characteristics of surveillance systems, and actual field conditions (eg, healthcare capacity) including the availability and effectiveness of the interventions, public health experts execute a large number of

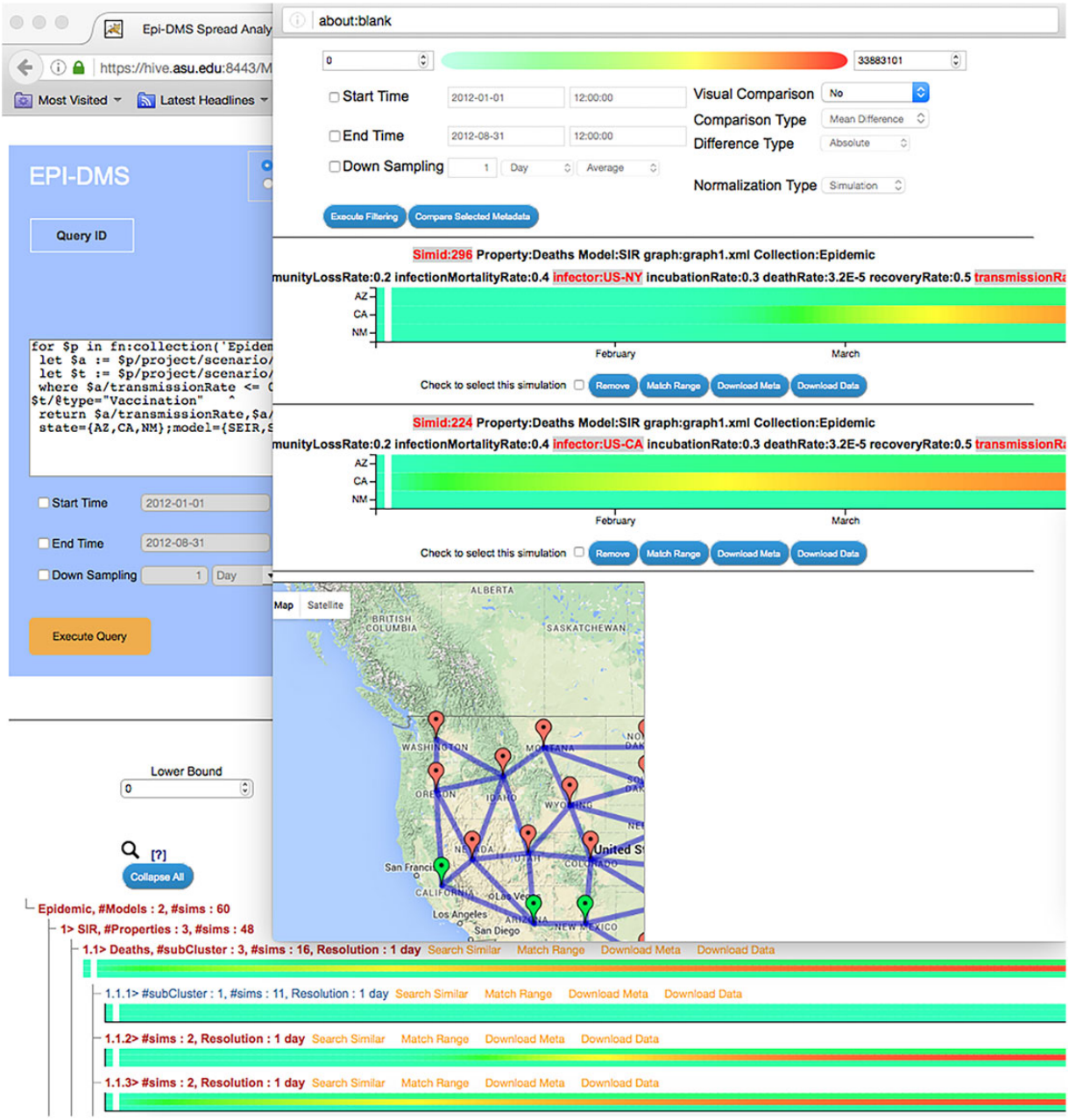

Figure 2. A sample epiDMS screenshot, which includes scenario-based querying and exploration. The figure shows a query posed to epiDMS, the set of results (visualized in the form of a navigable hierarchy of heat maps), and 2 simulations selected for detailed comparison. Please see the accompanying Supplementary Materials and the video available at https://www.youtube.com/watch?v=9w-4nDhXv3k for more details.

simulations by using the epiRun simulation ensemble creation engine to generate simulation instances. The configuration file for epiRun specifies applicable disease models, parameter value ranges and sampling granularities, connectivity and mobility graph assumptions, simulation duration, and assumptions regarding when and what interventions are to be applied. Given these, epiRun schedules the execution of these simulations. The simulation metadata and results are then read and stored in epiStore. Intuitively, each simulation result corresponds to a so-called possible world, and thus it is annotated and indexed with the metadata describing the corresponding scenario. Later, during hypothetical public health planning or pandemic response, the simulation results stored in epiStore can be accessed through scenario-based or observational search.

Scenario-Based Querying and Exploration

A basic functionality of epiDMS is to retrieve epidemic simulations, stored in epiStore, based on a user-specified scenario

description. For example, the user can formulate a query that asks the system to identify all preexecuted simulations, based on susceptible-exposed-infectious-removed and susceptible-infectious-removed epidemic models, where the input transmission rate parameter is set between 0.3 and 0.6 per day, the recovery rate parameter is set to 0.5 per day, and a vaccination-type trigger was used in the simulation. The query also specifies a particular mobility graph, describing expected movements of the populations during the epidemic, as an underlying assumption. In addition, the query asks the system to return daily (1-D) averages of infected, incidence, and deaths simulation output parameters for Arizona, California, and New Mexico for an epidemic simulation that lasts 8 months. Details of this query, as well as a detailed description of the query and visual exploration interface provided by epiDMS, are available in the Supplementary Materials.

Once the query is executed and the relevant simulations are identified, epiDMS then organizes the results in the form of a navigable hierarchy, based on the temporal dynamics of the disease: scenarios that result in similar patterns are grouped under the same branch, while simulations that show key differences in disease development are placed under different branches of the navigation hierarchy. The user can then navigate on this hierarchy using drill-down and roll-up operations and filter sets of simulations for further analysis.

Observational Alignment Based Querying and Exploration

In addition to scenario-based filtering, search, and exploration, epiDMS also enables searching particular temporal patterns on the epidemic ensembles. During an epidemic, this feature allows the expert to identify a relevant subset of stored simulations that match actual disease patterns or specific targets for intervention measures. This facilitates public health decision-makers to identify the relevant parameters that characterize transmission patterns in near real time, forecast epidemic spread as the epidemic evolves, and assess the potential impact of intervention scenarios. This platform also allows the user to perform simulation refinements by narrowing the parameter space of possible worlds on the basis of the current state of the epidemic. Hence, the user can use epiDMS to run additional simulations within the constrained parameter space to obtain more-detailed simulations, possibly with additional intervention assumptions, that are relevant to the current state of the epidemic.

CONCLUSIONS

In this article, we have described and illustrated with an example epiDMS [26], a novel epidemic simulation data management system that supports the generation, search, visualization, and analysis, in a scalable manner, of large volumes of epidemic simulation ensembles for decision-making. The system aims to assist experts and decision-makers in exploring large epidemic simulation ensemble data sets through efficient metadata- and similarity-based querying, data analysis, and visual exploration.

Supplementary Data

Supplementary materials are available at http://jid.oxfordjournals.org. Consisting of data provided by the author to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the author, so questions or comments should be addressed to the author.

Notes

Acknowledgments. We thank the members of the EmitLab at Arizona State University for their contributions to epiDMS.

Financial support. This work was supported by the National Science Foundation (grants 1318788 and 1518939).

Potential conflicts of interest. All authors: No reported conflicts. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Chowell G, Fenimore PW, Castillo-Garsow MA, Castillo-Chavez C. SARS outbreaks in Ontario, Hong Kong and Singapore: the role of diagnosis and isolation as a control mechanism. J Theor Biol **2003**; 224:1–8.
2. Siu A, Wong YCR. Economic impact of SARS: the case of Hong Kong. Cambridge, MA: MIT Press, **2004**; 3:62–83.
3. Khan K, Arino J, Hu W, et al. Spread of a novel influenza A (H1N1) virus via global airline transportation. N Engl J Med **2009**; 361:212–4.
4. McKibbin WJ. The swine flu outbreak and its global economic impact. Brookings, **2009**. http://www.brookings.edu/research/interviews/2009/05/04-swine-flu-mckibbin. Accessed 10 May 2016.
5. Abubakar I, Gautret P, Brunette GW, et al. Global perspectives for prevention of infectious diseases associated with mass gatherings. Lancet Infect Dis **2012**; 12:66–74.
6. Anderson RM, May RM. Infectious diseases of humans. Oxford: Oxford University Press, **1991**.
7. Nishiura H, Castillo-Chavez C, Safan M, Chowell G. Transmission potential of the new influenza A(H1N1) virus and its age-specificity in Japan. Euro Surveill **2009**; 14:pii:19227.
8. Merler S, Ajelli M, Pugliese A, Ferguson NM. Determinants of the spatiotemporal dynamics of the 2009 H1N1 pandemic in Europe: implications for real-time modelling. PLoS Comput Biol **2011**; 7:e1002205.
9. Van den Broeck W, Gioannini C, Gonçalves B, Quaggiotto M, Colizza V, Vespignani A. The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. BMC Infect Dis **2011**; 11:37.
10. Colizza V, Barrat A, Barthelemy M, Valleron AJ, Vespignani A. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. PLoS Med **2007**; 4:e13.
11. Hollingsworth TD, Ferguson NM, Anderson RM. Will travel restrictions control the international spread of pandemic influenza? Nat Med **2006**; 12:497–9.
12. Scalia Tomba G, Wallinga J. A simple explanation for the low impact of border control as a countermeasure to the spread of an infectious disease. Math Biosci **2008**; 214:70–2.
13. Cauchemez S, Ferguson NM, Wachtel C, et al. Closure of schools during an influenza pandemic. Lancet Infect Dis **2009**; 9:473–81.
14. Centers for Disease Control and Prevention (CDC). Interim pre-pandemic planning guidance: community strategy for pandemic influenza mitigation in the United States—early, targeted, layered use of nonpharmaceutical interventions. Atlanta, GA: CDC, **2007**.
15. Wu JT, Cowling BJ, Lau EH, et al. School closure and mitigation of pandemic (H1N1) 2009, Hong Kong. Emerg Infect Dis **2010**; 16:538–41.
16. Longini IM Jr, Nizam A, Xu S, et al. Containing pandemic influenza at the source. Science **2005**; 309:1083–7.
17. STEM. The spatiotemporal epidemiological modeler project. http://www.eclipse.org/stem. Accessed 10 May 2016.
18. Socioeconomic Data and Applications Center. http://sedac.ciesin.columbia.edu. Accessed 10 May 2016.
19. Germann TC, Kadau K, Longini IM Jr, Macken CA. Mitigation strategies for pandemic influenza in the United States. Proc Natl Acad Sci USA **2006**; 103:5935–40.
20. Mossong J, Hens N, Jit M, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med **2008**; 5:e74.
21. Flahault A, Vergu E, Boelle PY. Potential for a global dynamic of Influenza A (H1N1). BMC Infect Dis **2009**; 9:129.
22. Kenah E, Chao DL, Matrajt L, Halloran ME, Longini IM Jr. The global transmission and control of influenza. PLoS One **2011**; 6:e19515.
23. Merler S, Ajelli M. The role of population heterogeneity and human mobility in the spread of pandemic influenza. Proc Biol Sci **2010**; 277:557–65.
24. Barrett CL, Eubank SG, Smith JP. If smallpox strikes Portland. Sci Am **2005**; 292:42–9.

25. Chao DL, Halloran ME, Obenchain VJ, Longini IM Jr. FluTE, a publicly available stochastic influenza epidemic simulation model. PLoS Comput Biol **2010**; 6:e1000656.
26. Epidemic Simulation Data Management System (epiDMS). https://hive.asu.edu:8443/MVTSDB/?p=epidemic. Accessed 10 May 2016.
27. Berkeley Madonna: modeling and analysis of dynamic systems. http://www.berkeleymadonna.com/. Accessed 10 July 2016.
28. Liu S, Garg Y, Candan KS, Sapino ML, Chowell G. NOTES2: Networks-Of-Traces for Epidemic Spread Simulations. In: 29th AAAI Conference on Artificial Intelligence, AAAI 2015 - Austin, United States. AI Access Foundation, **2015**:79–83.
29. Schifanella C, Candan KS, Sapino ML. Multiresolution tensor decompositions with mode hierarchies. Article 10. ACM Trans Knowl Discov Data **2014**; 8.
30. Wang X, Candan KS, Sapino ML. Leveraging metadata for identifying local, robust multi-variate temporal (RMT) features. In: IEEE 30th International Conference on Data Engineering. Chicago, IL: IEEE, **2014**:388–99.