EPIDMS: DATA MANAGEMENT AND ANALYTICS FOR DECISION MAKING FROM EPIDEMIC SPREAD SIMULATION ENSEMBLES[i,ii]

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

[1] School of Informatics, and Decision Systems Engineering, Arizona State University, Tempe, Arizona, 85287, USA

[2] Computer Science Department, University of Torino, Torino, 10149, Italy

[3] School of Public Health, Georgia State University, Atlanta, GA, 30033, USA

Corresponding author contact information:
K. Selcuk Candan
Professor, Computer Science and Engineering
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
P.O. Box 878809
Tempe, AZ 85287-8809
e-mail: candan@asu.edu
phone: (480) 965-2770
fax: (480) 965-2751

Alternate author contact information:
Sicong Liu
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
P.O. Box 878809
Tempe, AZ 85287-8809
e-mail: sliu104@asu.edu

Running title: EpiDMS: An epidemic simulation data management system

Abstract word count: 169

Manuscript word count: 2818

Article type: Major Article

Abstract

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 assessing the likely impact of different intervention strategies. However, the management and analysis of simulation ensembles stemming from large-scale computational models poses challenges particularly when dealing with multiple inter-dependent parameters, spanning multiple layers and geo-spatial frames, affected by complex dynamic processes operating at different resolutions.

Methods: We describe and illustrate with examples a novel epidemic simulation data management system which 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 conclusion: 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 health-care emergencies and enabling critical services with significant economic and health impact.

Keywords: Epidemics, big data, simulation ensembles, data management, analytics, public-health decision making.

1 Introduction

The potential for pandemics to rapidly generate morbidity, mortality, and economic impact 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 (Severe Acute Respiratory Syndrome) emergency, which originated in China and spread to 29 countries, generated important nosocomial outbreaks in several regions by August 2003 [8,24]. More recently, the 2009 A/H1N1 influenza pandemic originating in Mexico rapidly spread around the globe via the airline network and reached 20 countries with highest volume of passengers arriving from Mexico within a few weeks of epidemic onset [14]. Importantly, the economic impact associated with a pandemic similar to the 2009 A/H1N1 influenza pandemic has been estimated to cost the global economy between $360 billion and $4 trillion [17] for the first year of virus circulation.

Large-scale computational transmission models of infectious disease spread are increasingly becoming part of the toolkit to carry out inferences on the spread and control of infectious diseases. Examples of real-time analyses of epidemics supported by large-scale transmission models include:

- estimating transmissibility of an epidemic disease, such as influenza [2,3,21],
- forecasting the spatio-temporal evolution of pandemics at different spatial scales [19,27],
- assessing the effect of travel controls during the early epidemic phase [9,12,22],
- predicting the effect of school closures in mitigating disease spread [5,6,29],
- assessing the impact of reactive vaccination strategies [16],

These analyses, however, require access to, integration, and analysis of models and large volumes of data, including datasets from diverse sources in order to parameterize demographic characteristics, contact networks, age-specific contact rates, mobility networks, and health-care and control interventions.

In this paper, 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 type of hypothetical analyses of epidemic scenarios in order to

address critical questions including: (a) Can we identify and classify key events (e.g., epidemic peak timing, likely epidemic duration) during an infectious disease outbreak from large simulation ensembles? (b) Can we compare and summarize a large number of epidemic simulations and observations under different epidemiological scenarios? (c) Can we discover latent relationships and dependencies among disease dynamics and social parameters?

1.1 *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 then may seed the infection in other regions, starting a new outbreak. Thus, large-scale epidemic simulation systems (e.g., GLEaM [27] and STEM [26]) 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 Socio-Economic Data and Applications Center (SEDAC) [25], which has a resolution of 15 × 15 minutes of arc.
- Mobility models can include long-range air travel mobility data, 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 has become available for a number of countries including the US [11], and South-East 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 [10,13,18].
- Epidemic models allow the user to specify epidemiological parameters that are specific of the infectious disease (such as transmissibility and seasonality), initial outbreak conditions (e.g.

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 infection, the parameters of a typical model (the modified Susceptible-Latent-Infectious-Recovered model described in [27]) include (a) the infection rate of contracting illness when an individual interacts with an infectious person; (b) infection rate scaling factors for asymptomatic infectors and treated infectors; (c) average length of the latency period (in which the individual is infected, but not infecting); (d) probability of symptomatic vs. asymptomatic infections; (e) change in the travelling behavior after the onset of symptoms; (f) average length of recovery; (g) percentage of infectious individuals that undergo pharmaceutical treatment; and (h) impact (e.g. on the length of the infectious period) of the treatment.

The output of a simulation is a multi-variate 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.

1.2 *Challenges*

While large-scale epidemic simulation systems such as GLEaM [27] or STEM [26] represent very powerful and highly modular and flexible epidemic spread simulation systems, their power for real-time decision making could be enhanced by addressing the following challenges:

(a) *Complexity of the simulation and observation data*. A sufficiently useful disease spreading simulation system requires models, including social contact networks, local and global mobility patterns of individuals, and epidemiological parameters for the infectious disease (e.g., infectious period). Epidemic simulations track 10s or 100s of inter-dependent parameters, spanning multiple layers and geo-spatial frames, affected by complex dynamic processes operating at different resolutions. Moreover, an ensemble of stochastic epidemic realizations may include 100s or 1000s of simulations, each with different parameters settings corresponding to slightly different, but plausible, scenarios [4,7]. As a consequence, running and interpreting simulation results (along with the real-world observations) to generate timely actionable results pose challenges.

(b) *Dynamicity of the real-world observations.* A major challenge in using data- and model-driven computer simulations for predicting geo-temporal evolution of epidemics for managing health

emergencies, such as the 2014-15 Ebola epidemic in West Africa, is that the data, models, and the 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: (a) revisions involve incorporating the real-world observations as well as updated probability surfaces into existing simulations to alter their outcomes; (b) 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.

In order 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 (a) executing large-scale simulation ensembles under a large number of diverse hypotheses/scenarios, and (b) 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 spatio-temporal spread at different spatial scales, 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.

2 EpiDMS System Overview and Use Scenario

The key characteristics of data and models relevant to data-intensive simulations include the following: (a) voluminous, (b) multi-variate, (c) multi-resolution, (d) multi-layer, (e) geo-temporal, (f) inter-connected

and inter-dependent, and (g) often incomplete/imprecise. Moreover, data and models dynamically evolve over time, due to control actions taken by individuals and public health interventions, requiring continuous adaptation and re-modeling.

The novel epiDMS software framework [1] 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 [30], the services provided by epiDMS include

- storage and indexing of large ensemble simulation data sets and the corresponding models; and
- search and analysis of ensemble simulation data sets to enable ensemble-based decision support [15,23,28].

The target user group for epiDMS include 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.

2.1 *System Overview*

The *epidemic simulation data management system* (epiDMS [1]) for managing the data and models for data-driven real-time epidemic simulations consists of three major components (Figure 1):

- *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 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 (i.e., time series for each output variable) then become inputs for the epidemic data and model store (epiStore), described next.

- *Epidemic data and model store* (epiStore) stores, and indexes the relevant data and metadata sets. The data and models relevant for modeling large-scale epidemics include the following:
    - Network layers: An epidemic simulation requires one or more layers of networks, from local and global mobility patterns to social contact networks.
    - Disease models, describing the epidemiological parameters relevant to a simulation and the parameter dependencies necessary in the computation of the disease spread.
    - Simulation time series: For a given disease study, researchers and decision makers often perform multiple simulations, each corresponding to different sets of assumptions (disease parameters or models) or context (e.g. spatio-temporal context, outbreak conditions, interventions).
    - Disease observations: These 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 (simulation model, parameter values, connectivity graphs) and simulation outputs (time series) and provides data analysis (such as clustering, classification, 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.
- *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.

2.2 *EpiDMS Use Scenario*

Let us 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 (e.g, healthcare capacity) including the availability and effectiveness of the interventions, public health experts execute a large number of simulations 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 "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.*

2.2.1 Scenario-based Querying and Exploration

A basic functionality of the epiDMS system 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 pre-executed simulations, based on SEIR (susceptible-exposed-infectious-removed) and SIR (susceptible-infectious-removed) epidemic models, where the input transmission rate parameter was set between 0.3 and 0.6, the recovery rate parameter was set to 0.5, 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 (AZ), California (CA), and New Mexico (NM), for an epidemic simulation that lasts 8 months (*Please see the online supplement for the details of this query as well as a detailed description of the query and visual exploration interface provided by epiDMS*).

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.

2.2.2 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 1) identify the relevant parameters that characterize transmission patterns in near real time, 2) forecast epidemic spread as the epidemic evolves, 3) assess potential impact of intervention scenarios. This platform also allows the user to perform simulation refinements by narrowing down the parameter space of “possible worlds” based on 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.

3 Conclusions

In this paper, we describe and illustrate with an example a novel epidemic simulation data management system (EpiDMS [1]) 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.

Acknowledgements

We thank the members of the EmitLab at ASU for their contributions to the epiDMS system. Please see the footnotes for the funding information and the conflict of interest statement.

4 References

1. Epidemic Simulation Data Management System (EpiDMS). Available at: https://hive.asu.edu:8443/MVTSDB/?p=epidemic. Accessed 10 May 2016.
2. Abubakar I, Gautret P, Brunette GW, Blumberg L, Johnson D, Poumerol G, et al. Global perspectives for prevention of infectious diseases associated with mass gatherings. Lancet Infect Dis. Jan;12(1):66-74.
3. Anderson RM, May RM. Infectious diseases of humans. Oxford: Oxford University Press; 1991.
4. Barrett CL, Eubank SG, Smith JP. If smallpox strikes Portland. Scientific American. 2005 Mar;292(3):42-9.
5. Cauchemez S, Ferguson NM, Wachtel C, Tegnell A, Saour G, Duncan B, et al. Closure of schools during an influenza pandemic. Lancet Infect Dis. 2009 Aug;9(8):473-81.
6. Centers for Disease Control and Prevention. Interim pre-pandemic planning guidance: community strategy for pandemic influenza mitigation in the United States—early, targeted, layered use of nonpharmaceutical interventions. Atlanta (GA): The Centers for Disease Control and Prevention; 2007.
7. Chao DL, Halloran ME, Obenchain VJ, Longini IM, Jr. FluTE, a publicly available stochastic influenza epidemic simulation model. PLoS Comput Biol. Jan;6(1):e1000656.
8. 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 Sep 7;224(1):1-8.
9. 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 Jan;4(1):e13.
10. Flahault A, Vergu E, Boelle PY. Potential for a global dynamic of Influenza A (H1N1). BMC Infect Dis. 2009;9:129.
11. Germann TC, Kadau K, Longini IM, Jr., Macken CA. Mitigation strategies for pandemic influenza in the United States. Proc Natl Acad Sci U S A. 2006 Apr 11;103(15):5935-40.
12. Hollingsworth TD, Ferguson NM, Anderson RM. Will travel restrictions control the international spread

of pandemic influenza? Nat Med. 2006 May;12(5):497-9.

13. Kenah E, Chao DL, Matrajt L, Halloran ME, Longini IM, Jr. The global transmission and control of influenza. PLoS One. 2011;6(5):e19515.
14. Khan K, Arino J, Hu W, Raposo P, Sears J, Calderon F, et al. Spread of a novel influenza A (H1N1) virus via global airline transportation. N Engl J Med. 2009 Jul 9;361(2):212-4.
15. Liu S., Garg Y, Candan KS, Sapino ML, Chowell G. NOTES2: Networks-Of-Traces for Epidemic Spread Simulations. AAAI Workshop on Computational Sustainability, 2015.
16. Longini IM, Jr., Nizam A, Xu S, Ungchusak K, Hanshaoworakul W, Cummings DA, et al. Containing pandemic influenza at the source. Science. 2005 Aug 12;309(5737):1083-7.
17. Warwick J. McKibbin. The Swine Flu Outbreak and its Global Economic Impact. Brookings. May 4th 2009. Available at http://www.brookings.edu/research/interviews/2009/05/04-swine-flu-mckibbin. Accessed May 10th 2016.
18. Merler S, Ajelli M. The role of population heterogeneity and human mobility in the spread of pandemic influenza. Proc Biol Sci. Feb 22;277(1681):557-65.
19. 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 Sep;7(9):e1002205.
20. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 2008 Mar 25;5(3):e74.
21. 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(22):pii: 19227.
22. 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 Jul-Aug;214(1-2):70-2.
23. Schifanella C, Candan KS, Sapino ML. Multiresolution Tensor Decompositions with Mode Hierarchies. ACM Trans. Knowl. Discov. Data (TKDD), 8(2), Article No. 10, June 2014.
24. Siu A and Wong Y C R. Economic Impact of SARS: The Case of Hong Kong. MIT Press. 2004, 3(1), p. 62-83.

25. Socioeconomic Data and Applications Center (SEDAC). Columbia University; Available at: http://sedac.ciesin.columbia.edu. Accessed 10 May 2016.
26. STEM. The spatiotemporal epidemiological modeler project. Available at http://www.eclipse.org/stem. Accessed 10 May 2016.
27. Van den Broeck W, Gioannini C, Goncalves 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.
28. Wang X, Candan KS, Sapino ML. Leveraging metadata for identifying local, robust multi-variate temporal (RMT) features. IEEE International Conference on Data Engineering (ICDE) 2014: 388-399.
29. Wu JT, Cowling BJ, Lau EH, Ip DK, Ho LM, Tsang T, et al. School closure and mitigation of pandemic (H1N1) 2009, Hong Kong. Emerg Infect Dis. 2010 Mar;16(3):538-41.
30. Berkeley Madonna - Modeling and Analysis of Dynamic Systems. http://www.berkeleymadonna.com/. Accessed 10 July 2016.

331 **Footnotes Page**

[i] Funding statement: This work is supported by NSF grants NSF # 1318788 and NSF # 1518939.

[ii] Conflict of interest statement: The conflict of interest (COI) disclosure forms are attached. Authors report no competing interests related to this manuscript.

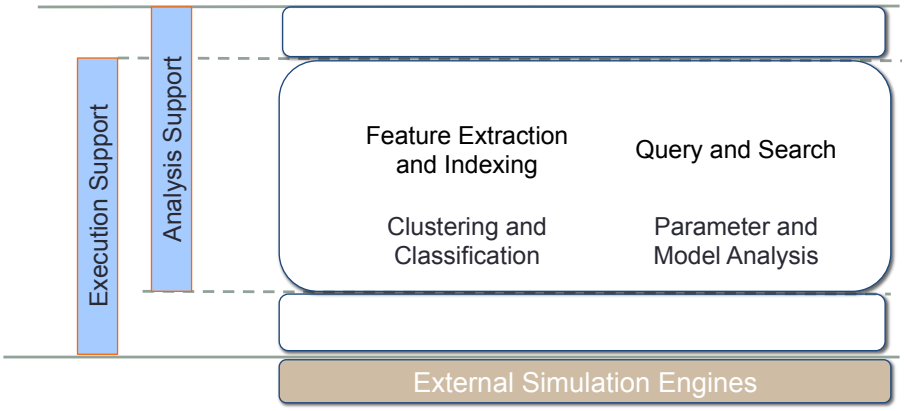


Figure 1. EpiDMS system overview

Figure

Click here to download Figure jid16_figure2_version2.docx

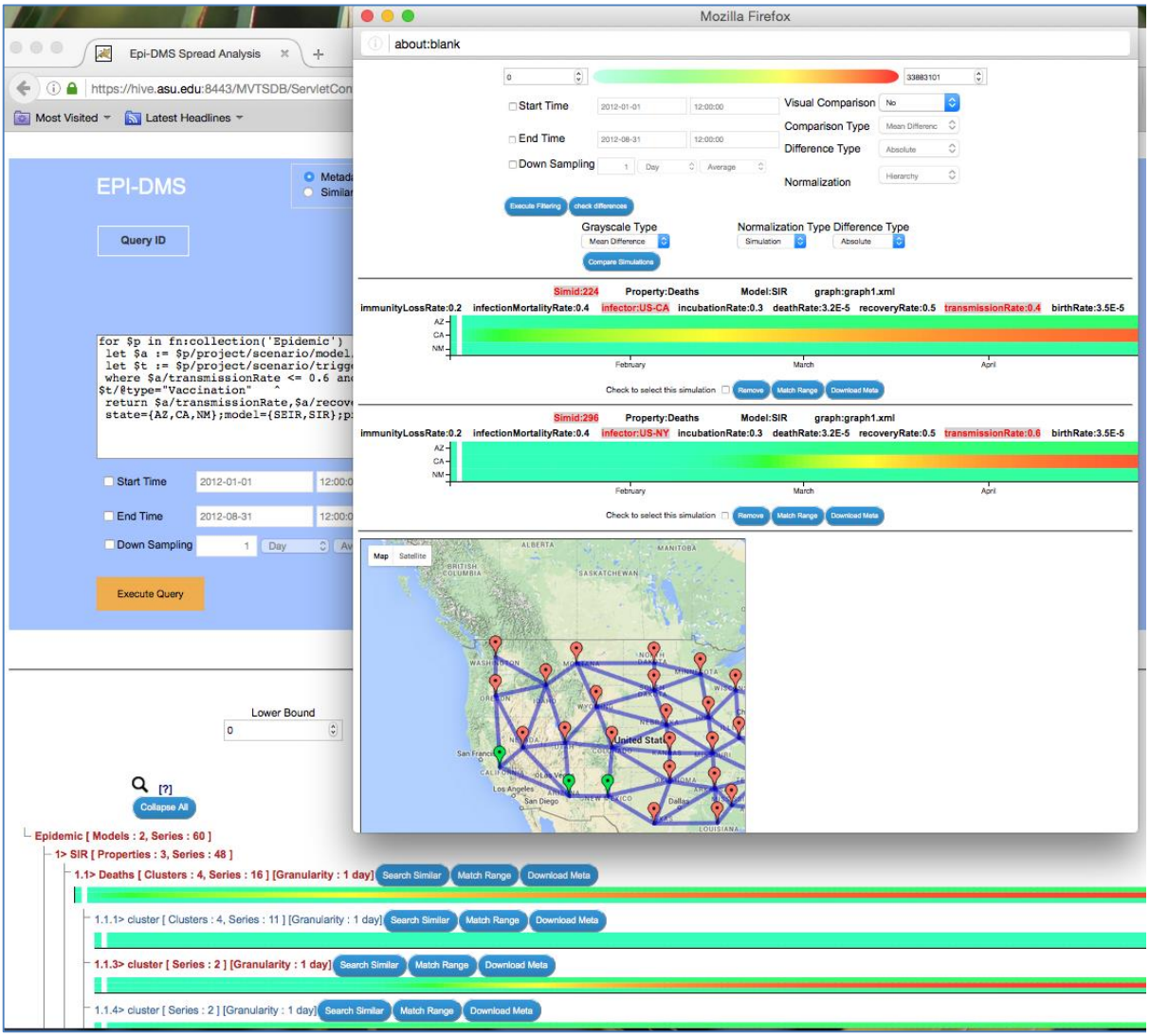


Figure 2. A sample epiDMS screenshot, which includes scenario-based querying and exploration: the figure shows a query posed to the epiDMS system, the set of results (visualized in the form of a navigable hierarchy of heatmaps) and two simulations selected for detailed comparison. Please see the accompanying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

EPIDMS: DATA MANAGEMENT AND ANALYTICS FOR DECISION MAKING FROM EPIDEMIC SPREAD SIMULATION ENSEMBLES

(ONLINE SUPPLEMENT)

Sicong Liu
Arizona State University
s.liu@asu.edu

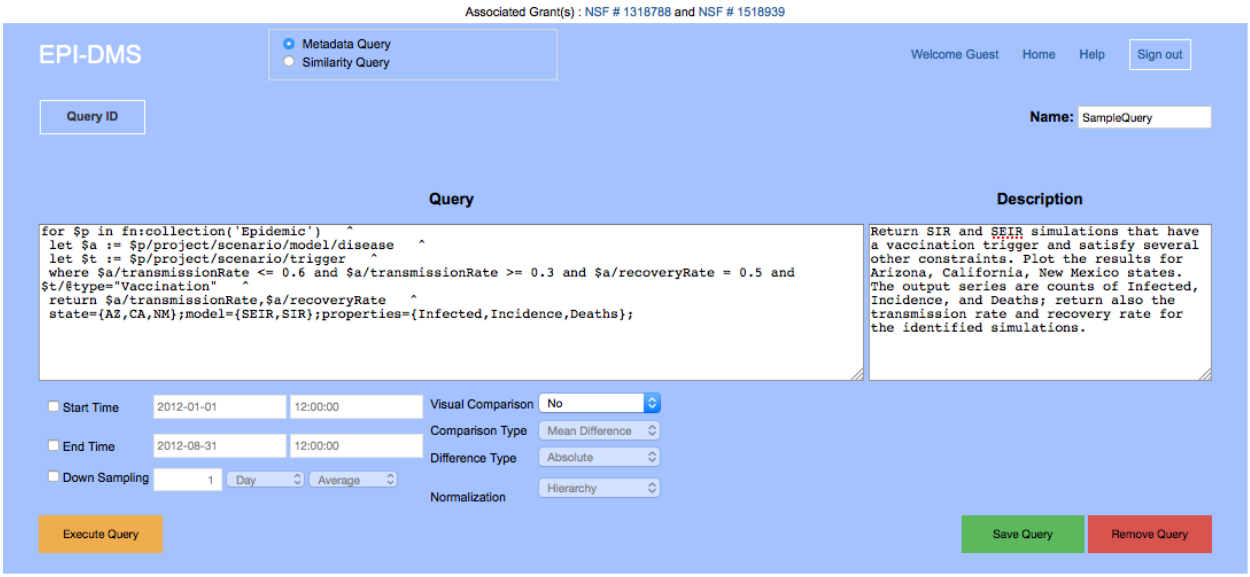
Silvestro Poccia
University of Torino
silvestro.poccia@edu.unito.it

K. Selcuk Candan
Arizona State University
candan@asu.edu

Gerardo Chowell
Georgia State University
gchowell@gsu.edu

Maria Luisa Sapino
University of Torino
marialuisa.sapino@unito.it

A basic functionality of the epiDMS system is to retrieve epidemic simulations, stored in epiStore, based on the user specified scenario description.



The basic query interface, visualized above, provides the following functionalities:

- Query Menu --- visualizes the list of queries that are stored in the system.
- Query Box --- visualizes the selected query and/or allows the user to edit a query

- Query Description --- shows the description of the selected query and/or allows the user to add a query description.

EpiDMS provides a rich query language to specify user queries. Consider, for example, the following sample query:

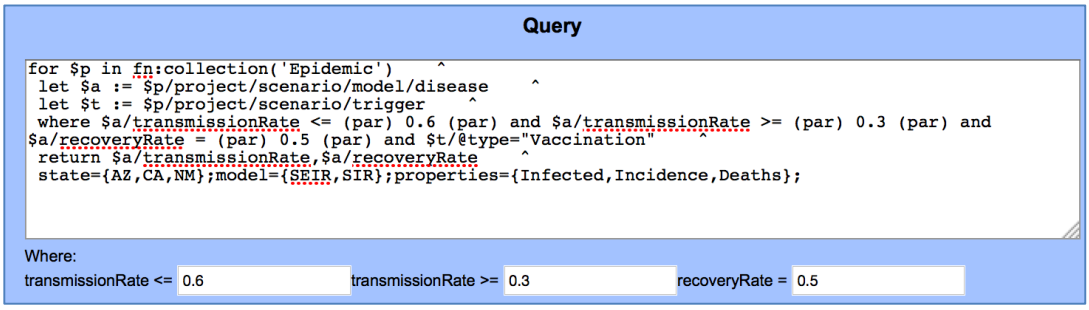
1. FOR $p in fn:collection('EpidemicSimulationEnsemble') ^
2. LET $diseaseModel := $p/project/scenario/model/disease ^
    - LET $triggerModel := $p/project/scenario/trigger ^
    - LET $epidemicScenario := $p/project/scenario ^
3. WHERE
    a. $diseaseModel/transmissionRate <= 0.6 and
    b. $diseaseModel/transmissionRate >= 0.3 and
    c. $diseaseModel/recoveryRate = 0.5 and
    d. $triggerModel/@type="Vaccination" and
    e. ($epidemicScenario/infector/@targetISOKey="US-CA" or $epidemicScenario/infector/@targetISOKey="US-NY") and
    f. ($epidemicScenario/graph = "mobility_graph_7.xml" or $epidemicScenario/graph = "mobility_graph_8.xml") ^
4. RETURN
    a. $diseaseModel/transmissionRate,
    b. $diseaseModel/recoveryRate,
    c. $epidemicScenario/graph ^
    d. STATE={AZ,CA,NM};
    e. MODEL={SEIR,SIR};
    f. PROPERTIES={Infected,Incidence,Deaths};
5. FROM ={01/01/2012 12:00:00}; TO={08/31/2012 12:00:00};
6. BY={1-D}; FUNCTION ={avg};

We describe the different components of this sample query below:

1. The “FOR” statement allows the user select the simulation dataset to query. In this example, the user selects to focus on the stored simulation set “EpidemicSimulationEnsemble”.
2. The “LET” statement allows to associate variables representing disease and intervention trigger models and epidemic scenarios.
3. The “WHERE” clause allows the user to specify conditions on the simulation models to filter those simulations that are relevant for the current analysis. In this example, the user specifies that for the returned simulations, the transmission rate parameter should be between 0.3 and 0.6, the recovery rate parameter should be set to 0.5, and that a “vaccination” type trigger should be included in the simulation model. The user also specifies that epidemic should have started at California (CA) or New York (NY) state and the “mobility_graph_7.xml” or “mobility_graph_8.xml” should have been used to generate the simulations.
4. The “RETURN” clause lists the simulation parameters to be returned in the result. In this example, the user is interested in the transmission rate, recovery rate, the mobility graph for each returned simulation. In addition, the query asks the system to return the time series corresponding to the “infected”, “incidence”, and “deaths” simulation output parameters for Arizona (AZ), California (CA), and New Mexico (NM) states.
5. In this clause, the user specifies that s/he is interested in only the first 8 months of the simulation.
6. Furthermore, the user specifies that the system return daily (1-D) averages of the simulation parameters for the specified duration.

1.1.1 Query Interface

The epiDMS query interface allows the user to specify and execute parametric queries. As illustrated below, parametric queries support query specification reuse – instead of writing a new query for different parameters, the user can specify and store a parametric query, which can then be invoked with different parameter values, as seen in the following example:

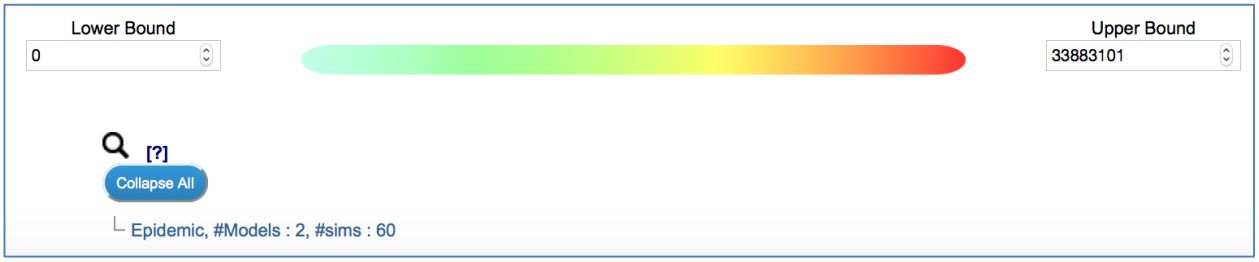


In the above example, those query parameters whose values are bracketed with the symbol "(par)" are interpreted as being parametric. The user can vary these values using a form-based interface without having to modify the source code directly.

1.1.2 Result Set Exploration Module

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 on this hierarchy and pick sets of simulations to study and compare in further detail the corresponding scenarios. This process is described below:

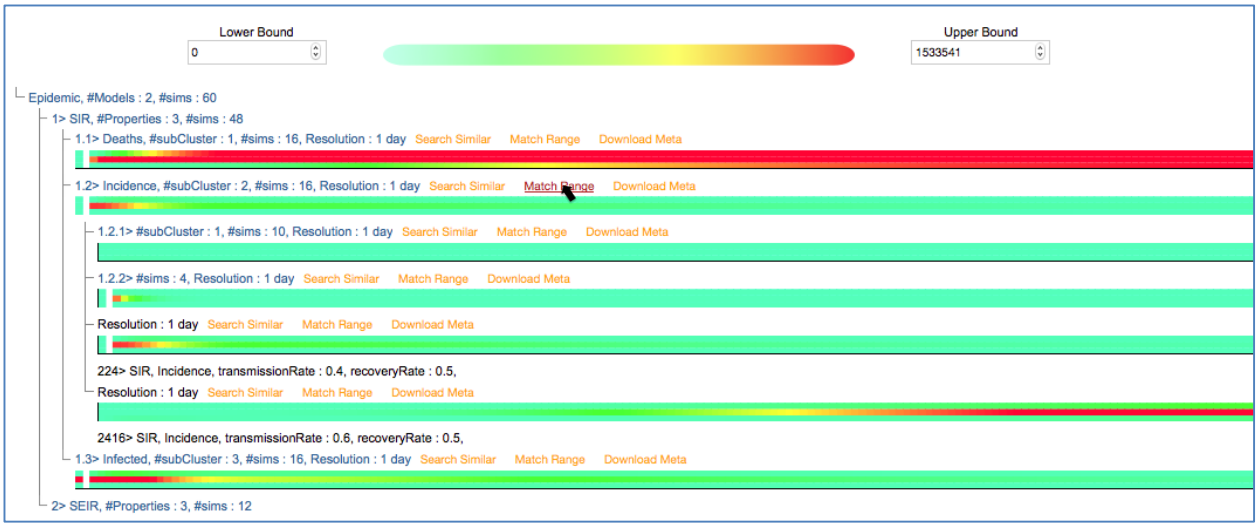
Once the matching simulations are identified, the user is presented with an initially collapsed hierarchy of results:



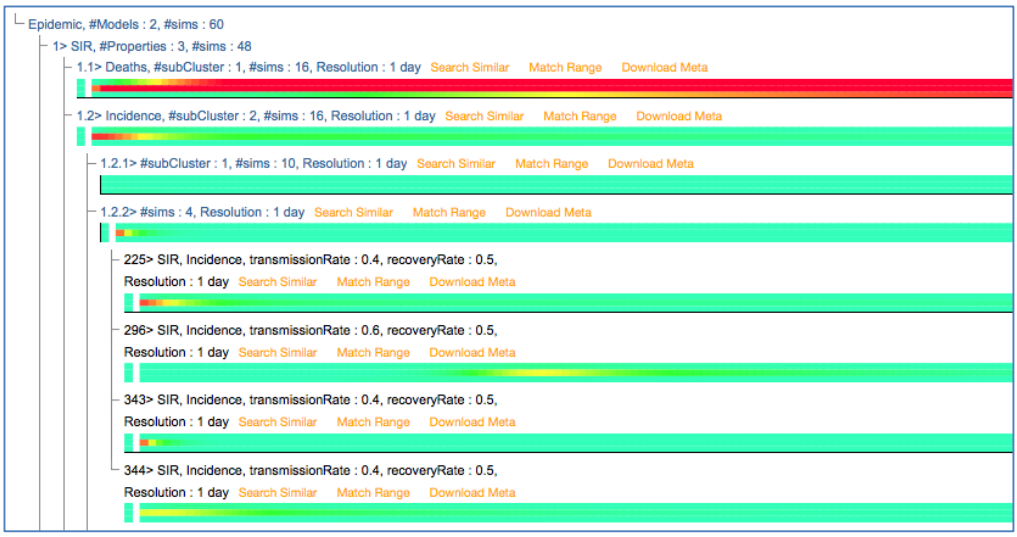
Above, we see that the query identified 60 matching simulations, from two different disease models. The legend at the top provides the scope of values in the results. The user can explore these simulations by drilling down or rolling up on the result hierarchy:

As we see above, the top-level of the result hierarchy includes the disease models (SIR and SEIR in this example). At the next level, the user is presented the output parameters specified in the query ("incidence", "deaths", and "infected" in this example). Under this level, the results are organized in the form of a cluster hierarchy, where similar simulations are clustered under the same navigation branch. For each node in the navigation hierarchy, a cluster representative is selected and the corresponding simulation is visualized in the form of a heatmap, where each row corresponds to a location (states "AZ". "CA", and "NM") in this example. The user can obtain detailed information about the presented simulations, by hovering the mouse on the heatmaps or download simulation results (in the form of CSV files) or metadata and model specifications (in XML format) corresponding to different simulations for further study or dissemination to decision makers.

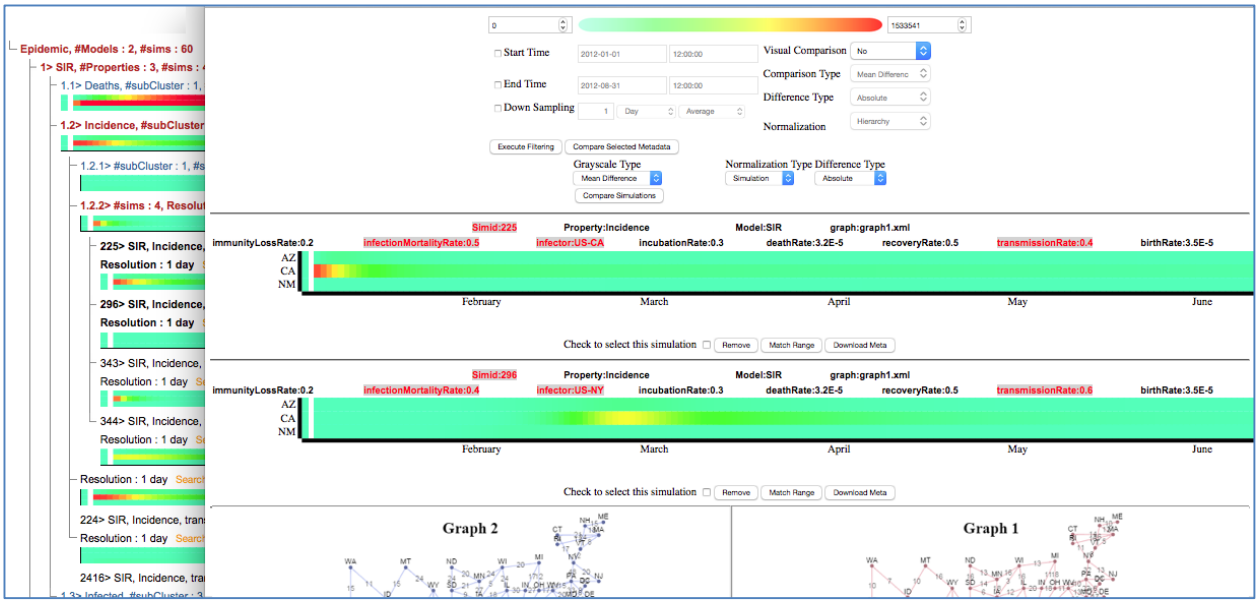
The user can also use the "match range" feature to change the scale of visualization so that the upper bound of visualized values in the heatmap is modified in a way that matches a selected simulation to enable better visualization of its details. For example, in the example below, the heatmap scale has been modified to match the number of incidences, rather than the number of deaths; thus, we are able to better observe the differences among the incidence clusters:



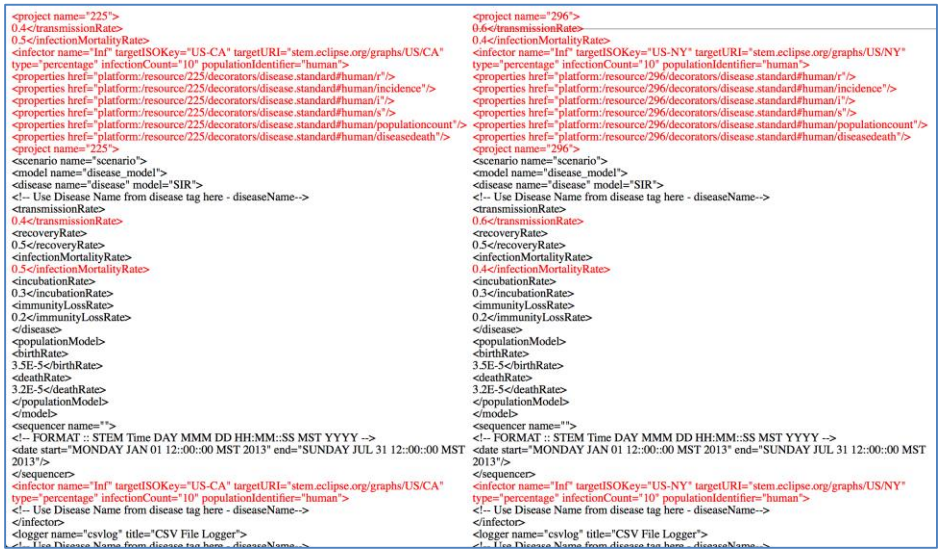
The user can explore these simulations by navigating on the hierarchy by drilling down or rolling up different branches. In the following example, the user has drilled down on the cluster 1.2.2 of the "incidence" data to observe the simulations clustered under this navigation node:



To further study individual simulations, the user then can double click on any simulation in the navigation hierarchy to place them in to a separate comparison interface. In the example, shown below, the user selected two simulations (#225 and #296) under cluster 1.2.2 to be studied and compared further in detail:



In the detailed comparison interface, the user can compare the two (or more) simulations side by side and observe the differences in the input parameters and models. The user can further ask the system to visualize the precise differences in the metadata corresponding to selected pairs of simulations:



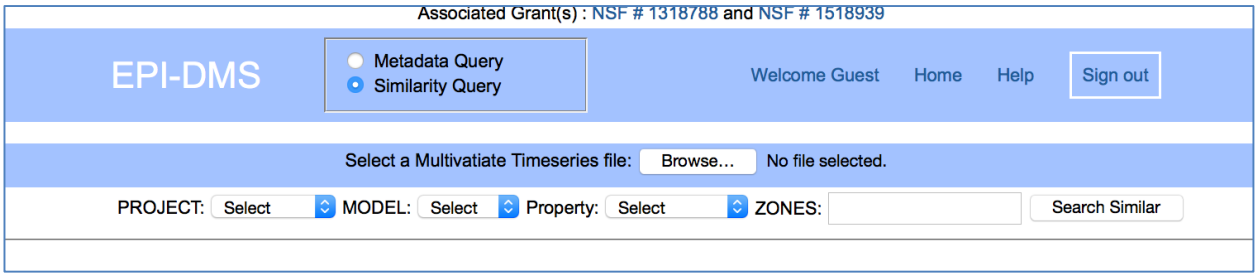
Here the text highlighted in red points to the differences in metadata corresponding to the pair of simulations selected or comparison.

1.1.3 Observational Similarity 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.

To use similarity based querying, the user can either click on the “Search Similar” option on the result visualization interface or switch to the “Similarity Query” interface and provide a file which contains the observations of interest:



Once a simulation/observation and states of interest are provided, the system searches in the databases existing simulations that show a similar pattern. Results are ranked in terms of their similarities to the provided query pattern:

Note that, once again, the user can obtain detailed information about the presented simulations by

hovering the mouse on the heatmaps or download simulation data or metadata corresponding to different simulations for further study. Moreover, as before, to further study individual simulations, the user can double click on any simulation in the navigation hierarchy to place them in to the comparison interface. Please see the accompanying video at https://www.youtube.com/watch?v=9w-4nDhXv3k for more details.

Frequently Asked Questions

Question #1*: "It appears that epiDMS would be operated by those with at least moderate infectious disease modeling experience. Is it true that epiDMS requires programming skills by the operator (while there appears to be a GUI, there also appears to be a moderate amount of programming involved in operating this)."*

Answer: The target user group for epiDMS include a range of public health researchers and decision makers. While creation of models for ensemble simulations and formulating queries over ensembles simulations require moderate infectious disease modeling experience and familiarity with (not programming, but) declarative querying, epiDMS also provides parameterized queries and other interactive user interfaces to enable decision makers with minimal experience to explore large ensemble simulations.

Question #2: *"Can you give a pathogen-specific example of a public health emergency in which the data, models and underlying model parameters dynamically evolve over time requiring continuous analyses and interpretations of the incoming data and adaptation of the networks and models."*

Response: The 2014-15 Ebola epidemic in West Africa was an example of such an health emergency where the situation (what we new about the disease characteristics, available and implemented intervention strategies, population dynamics, and social interactions among and within effected populations) continuously changed as the epidemic evolved, requiring reassessment and revisions models and re-interpretations of the data.

Question #3: *"How does epiDMS differ from existing modeling platforms and packages (e.g,. Berkeley Madonna or R)."*

Answer: Unlike other dynamic modeling platforms such as Berkeley Madonna, the services provided by epiDMS include

- storage and indexing of large ensemble simulation data sets and the corresponding and models; and
- search and analysis of ensemble simulation data sets to support ensemble-based decision support.

In that sense, epiDMS is less of a modeling tool and more of a multi-model, multi-instance ensemble simulation-based decision support system.

Question #4: *"Is epiDMS specific to a particular disease model or simulation engine? If not, how does different models fit within the database?"*

Response: We thank the reviewer for bringing to our attention that the original manuscript did not make it sufficiently clear that epiDMS is a model independent system by design:

- epiRun, for execution ensemble simulations, is not specific to any disease model or simulation engine and can wrap –as a black-box software component– any epidemic simulation engine as long as it provides command line invocation.
- epiStore, which stores epidemic models and the generated simulation ensembles, 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 is available. This wrapper based design ensures that models and simulations generated by different engines and tools can be imported into epiStore and queried and analyzed simultaneously irrespective of their origin.
- Finally, epiViz, which provides a web-based query and result visualization interface to support

user interaction and exploratory decision making is also model independent. More specifically, the underlying query specification language can support queries based on any model, without having to make any a priori assumptions regarding what the input and output parameters of the simulations are. Once they are imported into epiStore, parameters of any model can be queried, visualized, and explored.

The current alpha version of the system provides wrappers for the STEM simulation engine and can import models and simulations generated by STEM tool. The beta version of the tool will include wrappers for other systems.

Question #5: *"(i) What are the computational demands of epiDMS. e.g., can this be run on a standard laptop? A tablet/smartphone? From the video, it appears this is a web-based platform, but is there a stand alone downloadable form which can be run in potential areas with no internet connection (e.g., in certain field settings)?"*

Answer: The user interface of epiDMS is indeed a web-based platform and can run on any networked laptop and most tablets or smartphones. The backend, however, runs on server hardware. It is, however, possible to configure a laptop to act both as the backend and frontend.

Question #6: *"What is the speed of the simulation analyses?"*

Answer: This depends on the size of the simulation ensemble, number of variates/parameters of interest, the type of analysis, and the hardware configuration (memory, number of cores) at the back-end server platform. Having said that, we are doing our best to provide a near real-time and interactive experience to the users.

Question #7: *"What is the format of the modelling output? Can it easily be downloaded and disseminated to decision makers in public health practice?"*

Answer: Users of epiDMS can download simulation results (in the form of CSV files) or metadata and model specifications (in XML format) corresponding to different simulations for further study or

dissemination to decision makers.

Question #8: *"Can you confirm if this is a free system?...is there an open-source version of the software with scope for a community of developers?"*

Answer: An alpha version of the source-code for epiDMS is currently available upon request, and free of charge, to researchers and educators in the non-profit sector, including institutions of education, research, and government laboratories under an Apache 2.0 license (http://www.apache.org/licenses/LICENSE-2.0). The terms of the license allows individuals to modify the source code and to share modifications and also enable open source development of the software by other individuals and teams. The terms of software availability permits the commercialization of enhanced and customized versions of the software and incorporation of the software or pieces of it into other software packages. The beta release of the source-code will be available to the public through GitHub under the same terms.

Question #9: *"Is there a user-group forum for users to ask questions, trouble-shoot, show applications etc.?"*

Answer: While such a user-group forum does not currently exist, we will bootstrap a group along with the beta release of the system. In addition, we are planning to

- carry out demonstrations of epiDMS,
- give tutorials, and
- organize workshops

at leading forums targeting public healthcare researchers, scientists, and decision makers.

EPIDMS: DATA MANAGEMENT AND ANALYTICS FOR DECISION MAKING FROM EPIDEMIC SPREAD SIMULATION ENSEMBLES

Sicong Liu @ Arizona State University, USA
Silvestro Poccia @ University of Torino, Italy
K. Selcuk Candan @ Arizona State University, USA
Gerardo Chowell @ Georgia State University & National Institutes of Health, USA
Maria Luisa Sapino @ Arizona State University, USA & University of Torino, Italy

UNIVERSITÀ DEGLI STUDI DI TORINO

Supported by
- NSF III#1318788 “Data Management for Real-Time Data Driven Epidemic Spread Simulations”
- NSF RAPID “Understanding the Evolution Patterns of the Ebola Outbreak in West-Africa and Supporting Real-Time Decision Making and Hypothesis Testing through Large Scale Simulations”

Epidemics….

- SARS (Severe Acute Respiratory Syndrome) epidemic is estimated to have started in China in November 2002, had spread to 29 countries by August 2003
- A pandemic similar to the swine flu in 2009 is estimated to cost $360 billion in a mild scenario to the global economy and up to $4 trillion in an ultra scenario, within the first year of the outbreak
- The World Health Organization declared the Ebola epidemic in West Africa a *Public Health Emergency of International Concern* on August 8th, 2014, with exponential dynamics characterizing the initial growth in numbers of new cases in some areas

Epidemics….

- Data- and model-driven computer simulations are increasingly critical in predicting geo-temporal evolution of epidemics
  - estimating transmissibility of an epidemic disease, such as influenza,
  - forecasting the spatio-temporal spread of pandemic disease at different spatial scales,
  - assessing the effect of travel controls during the early stage of the pandemic,
  - predicting the effect of implementing school closures,
  - assessing the impact of pharmaceutical interventions on pandemic disease

Epidemics….

Not much room for error

Both action and inaction can have high costs in terms of their economic impacts and human lives affected

Critically needed…

- Tools that help
  - executing large-scale simulation ensembles under a large number of diverse hypotheses/scenarios, and
  - 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 spatio-temporal spread at different spatial scales,
    - assessing the cost and impact of interventions, including travel controls, at various stages of the epidemic
  - supporting real-time decision making and hypothesis testing through large scale simulations.

Good news: epidemic simulation software…

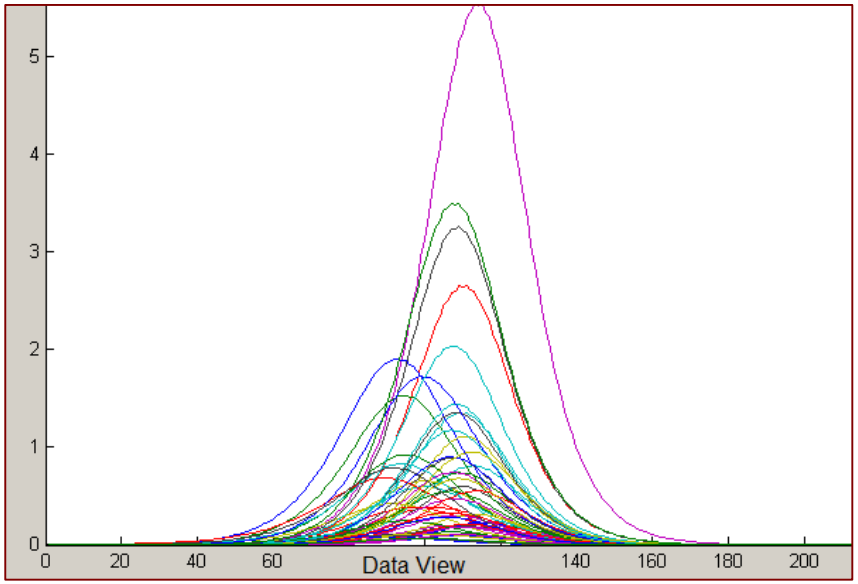
- Various time-step based epidemic spread simulation software exist (GLEaM, STEM)

Simulation model parameters…

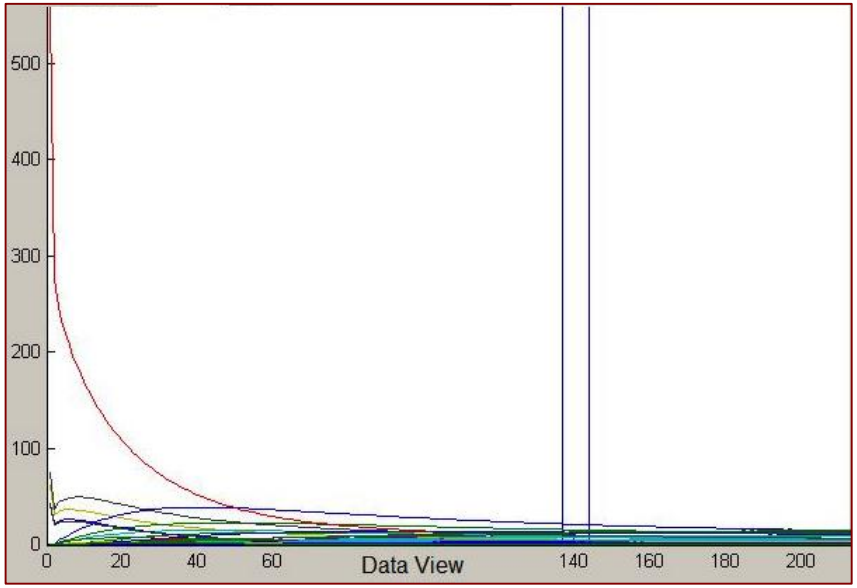
- Spatial/Demographic Layer
  - 3,362 subpopulations in 220 countries of the world)
- Mobility layer
  - long-range air travel mobility data, from the Inter. Air Transport Assoc. and the Official Airline
  - short-range commuting patterns between adjacent subpopulations
- Epidemic layer
    - infection rate of contracting illness when an individual interacts with an infectious person;
    - infection rate scaling factors for asymptomatic infectors and treated infectors;
    - probability of symptomatic vs. asymptomatic infections;
    - average length of the latency period (in which the individual is infected, but not infecting);
    - average length of recovery;
    - percentage of infectious individuals that undergo pharmaceutical treatment
    - impact of treatment (e.g. on the length of the infectious period)
    - change in the travelling behavior after the onset of symptoms;
  - Initial conditions of outbreak
  - intervention measures.

How do the simulation results look?

Each curve is a different US state



Simulation #1



Simulation #2

- These two simulation differ in
  - where the disease enters the US and
  - the disease characteristics, such as infection rate and recovery rate.

Bad news…

- Challenge #1: Epidemic simulations track
  - 100s of inter-dependent parameters,
  - spanning multiple layers and geo-spatial frames,
  - affected by complex dynamic processes operating at different resolutions.

- Challenge #2: Given the
  - unpredictability of an epidemic and
  - unpredictability of the actions of various independent agencies,

  decision makers need to generate many thousands of simulations, each with different parameters corresponding to plausible scenarios.

- Challenge #3: Simulations need to be continuously revised based on real-world data as the epidemic and intervention mechanisms evolve.

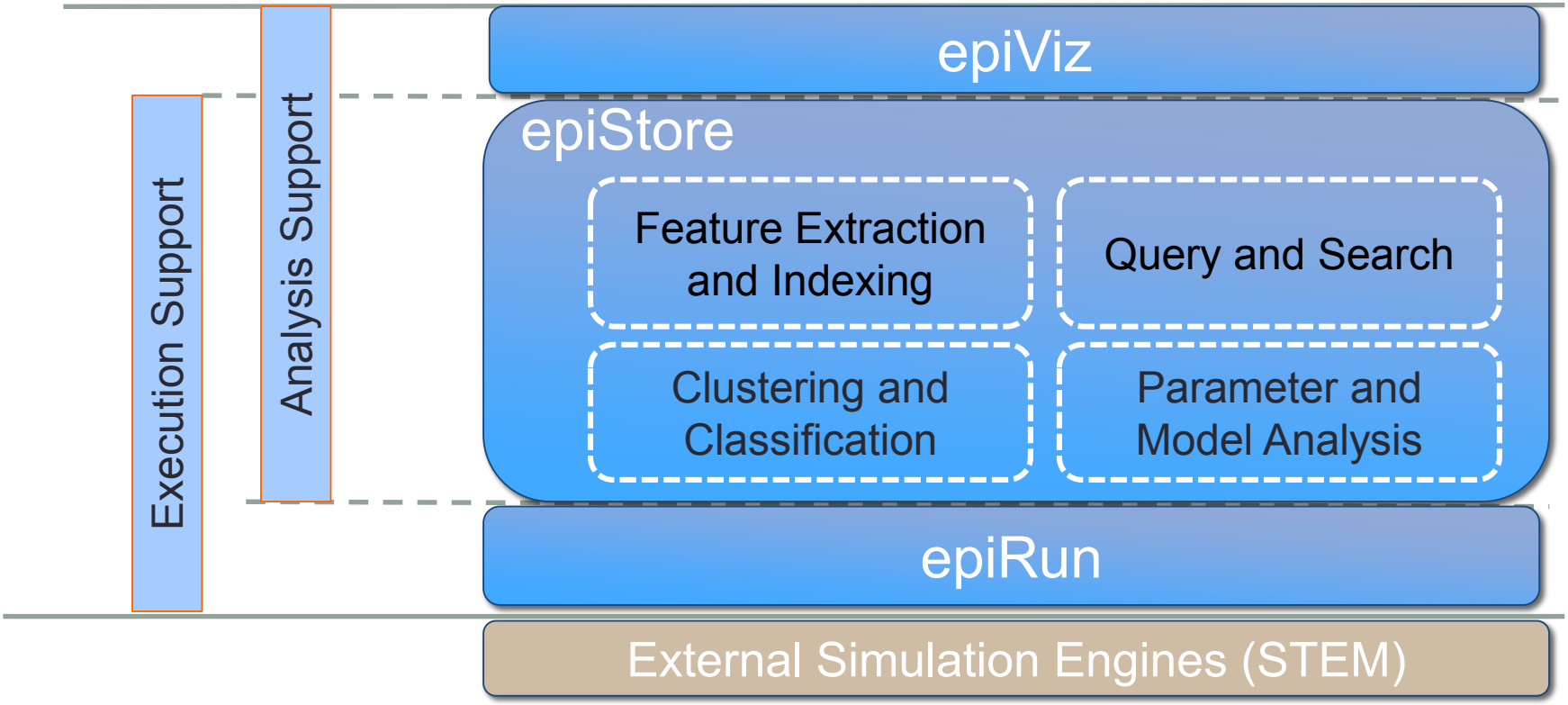
Challenges

- Because of the size and complexity of the data and the varying spatial and temporal scales at which the key processes operate; experts lack the means to
  - analyzing simulation results,
  - understanding relevant processes and
  - assessing the robustness of conclusions driven from the resulting simulations.

Questions (??).....

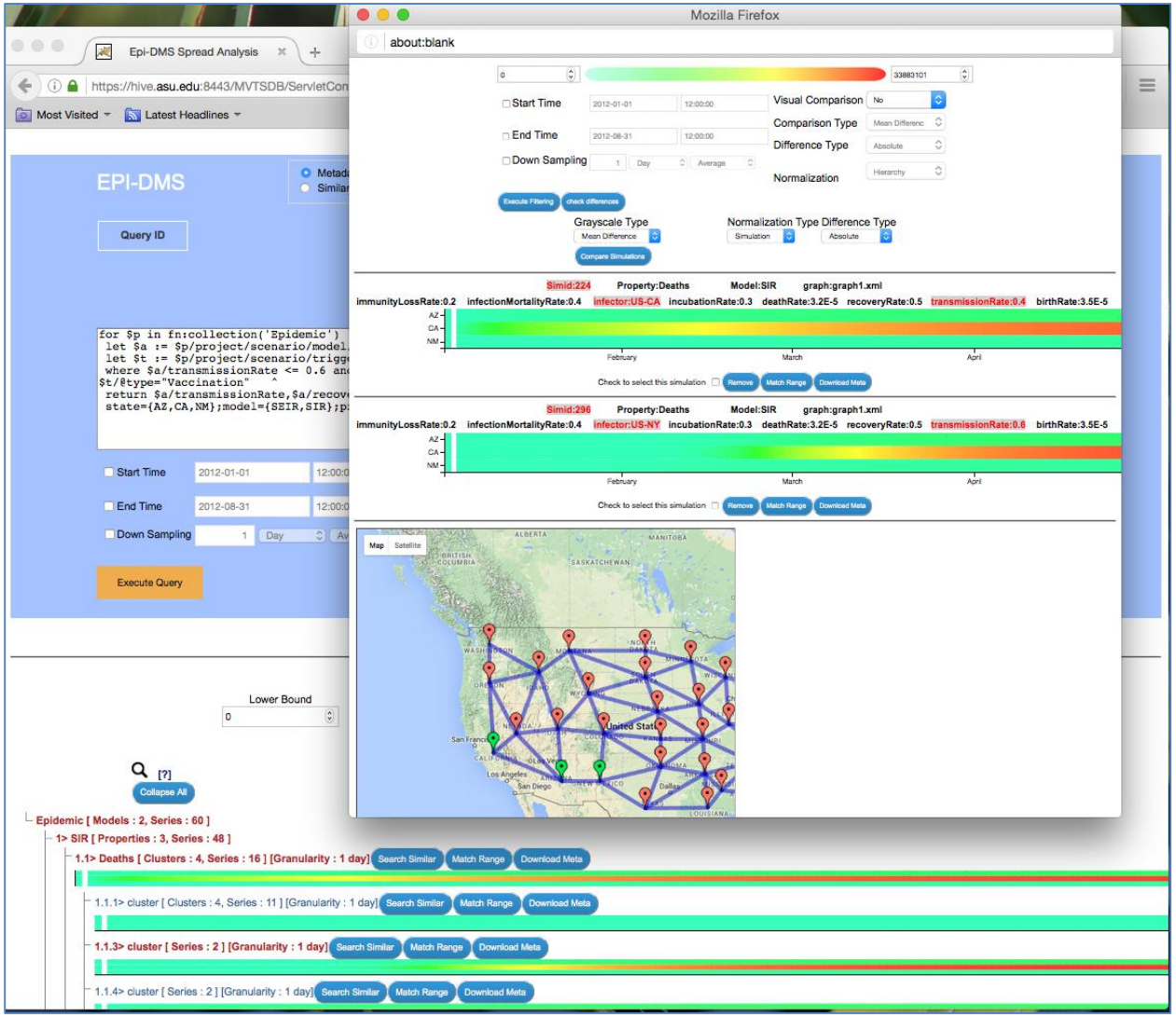
- Can we discover key events in a simulation trace and summarize a large simulation trace to highlight these key events?
- Can we classify these key events?
- Can we compare a large number of simulation traces and observations (under different parameter settings) to identify their similarities and differences?
- Can we analyze one or more simulation traces to discover underlying patterns and relationships between input parameters, key events/interventions, and simulation outcomes?
- Can we **search and retrieve simulation traces** based on the underlying key events or the overall trace similarities?

epiDMS Framework…



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.

EpiDMS Epidemic Simulation Ensemble Exploration Interface



epiDMS

- epiDMS facilitates public-health decision makers
  - identify the relevant parameters that characterize transmission characteristics,
  - forecast epidemic spread as the epidemic evolves,
  - assess potential impact of intervention scenarios.

- epiDMS also allows the user to
  - perform simulation refinements by narrowing down the parameter space based on the current state of the epidemic
  - run additional simulations within the new parameter space to obtain more detailed simulations relevant to the current disease state.

Conclusion

- A sample EpiDMS visualization interface is available at
  - http://aria.asu.edu/epidms
- You can also watch a tutorial at
  - https://www.youtube.com/watch?v=9w-4nDhXv3k
- For feedback, please contact:
  - candan@asu.edu

Supported by
- NSF III#1318788 "Data Management for Real-Time Data Driven Epidemic Spread Simulations"
- NSF RAPID "Understanding the Evolution Patterns of the Ebola Outbreak in West-Africa and Supporting Real-Time Decision Making and Hypothesis Testing through Large Scale Simulations"

Contributors..

- Reece Bailey, ASU
- K. Selcuk Candan, ASU
- Xilun Chen, ASU
- Gerardo Chowell-Puente, GSU
- Yash Garg, ASU
- Anisha Gupta, ASU
- Shengyu Huang, ASU
- Jung Hyun Kim, ASU
- Sicong Liu, ASU
- Sam Morton, ASU
- Parth Nagarkar, ASU
- Silvestro Poccia, U. Torino
- Rosaria Rossini, U. Torino
- Shivam Sadakar, ASU
- Maria Luisa Sapino, U. Torino
- Adam Tse, ASU
- Xiaolan Wang, ASU
- Fiona Zhang, ASU
- Luke Zhang, ASU

EPIDMS: DATA MANAGEMENT AND ANALYTICS FOR DECISION MAKING FROM EPIDEMIC SPREAD SIMULATION ENSEMBLES[i,ii]

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

[1] School of Informatics, and Decision Systems Engineering, Arizona State University, Tempe, Arizona, 85287, USA

[2] Computer Science Department, University of Torino, Torino, 10149, Italy

[3] School of Public Health, Georgia State University, Atlanta, GA, 30033, USA

Corresponding author contact information:
K. Selcuk Candan
Professor, Computer Science and Engineering
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
P.O. Box 878809
Tempe, AZ 85287-8809
e-mail: candan@asu.edu
phone: (480) 965-2770
fax: (480) 965-2751

Alternate author contact information:
Sicong Liu
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
P.O. Box 878809
Tempe, AZ 85287-8809
e-mail: sliu104@asu.edu

Formatted: Indent: Left: 0.5"

Running title: EpiDMS: An epidemic simulation data management system

Abstract word count: 169

Manuscript word count: ~~26172~~818

Article type: Major Article

36
37

Formatted: Justified, Right: -0.04", Space Before: 2.6 pt, Line spacing: 1.5 lines, No widow/orphan control, Font Alignment: Baseline

Formatted: Underline

Abstract

Background: Carefully calibrated large-scale computational models of epidemic spread represent a powerful tool to support the decision-making process during epidemic emergencies. ~~These~~Epidemic models are being increasingly used for generating forecasts of the spatial-temporal progression of epidemics at different spatial scales and assessing the likely impact of different intervention strategies. However, the management and analysis of simulation ensembles stemming from large-scale computational models poses challenges particularly when dealing with multiple inter-dependent parameters, spanning multiple layers and geo-spatial frames, affected by complex dynamic processes operating at different resolutions. *Methods*: We describe and illustrate with examples a novel epidemic simulation data management system which 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 conclusion: 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 health-care emergencies and enabling critical services with significant economic and health impact.

Keywords: Epidemics, big data, simulation ensembles, data management, analytics, public-health decision making.

1 Introduction

The potential for pandemics to rapidly generate morbidity, mortality, and economic impact 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 (Severe Acute Respiratory Syndrome) emergency ~~that, which~~ originated in China and spread to 29 countries ~~and,~~ generated important nosocomial outbreaks in several regions by August 2003 [~~25~~8,24]. More recently, the 2009 A/H1N1 influenza pandemic originating in Mexico rapidly spread around the globe via the airline network and reached 20 countries with highest volume of passengers arriving from Mexico within a few weeks of epidemic onset [14~~]. A~~]. Importantly, the economic impact associated with a pandemic similar to the 2009 A/H1N1 influenza pandemic has been estimated to cost ~~$360 billion in a mild scenario to~~ the global economy between $360 billion and ~~up to~~ $4 trillion ~~in an ultra scenario~~ [17~~], within~~] for the first year of virus circulation.

Large-scale computational transmission models of infectious disease spread are increasingly becoming part of the toolkit to carry out inferences on the spread and control of infectious diseases ~~[4].~~ Examples of real-time analyses of epidemics supported by large-scale transmission models include:

- estimating transmissibility of an epidemic disease, such as influenza [2,~~3~~3,21],
- ~~Estimating the risk of observing multiple generations of disease transmission in particular areas of the world~~
- forecasting the spatio-temporal evolution of pandemics at different spatial scales [19,27],
- assessing the effect of travel controls during the early epidemic phase [9,12,22],
- predicting the effect of school closures in mitigating disease spread [5,6,29],
- assessing the impact of reactive vaccination strategies [16],

These analyses, however, require access to, integration, and analysis of models and large volumes of data, including datasets from diverse sources in order to parameterize demographic ~~data~~characteristics, contact networks, age-specific contact rates, mobility networks, and ~~data on~~ health-care and control interventions ~~from diverse sources.~~

Formatted: Bold

In this paper, we argue that, if effectively leveraged, existing simulation analyses and real-time observations ~~incoming~~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 type of hypothetical analyses of epidemic scenarios in order to address critical questions including: (a) Can we ~~discover~~identify and classify key events ~~and summarize a~~(e.g., epidemic peak timing, likely epidemic duration) during an infectious disease outbreak from large simulation ~~trace to highlight these key events?~~ensembles? (b) Can we compare and summarize a large number of ~~simulation traces~~epidemic simulations and observations ~~(~~under different ~~parameter settings)?~~epidemiological scenarios? (c) Can we discover latent relationships and ~~structures~~dependencies among disease dynamics and social parameters?

1.1 *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 ~~local~~localized outbreak. These infected individuals then may seed the infection in other regions, starting a new outbreak. ~~Thus, state-of-the-art disease spread simulators, such as~~Thus, large-scale epidemic simulation systems (e.g., GLEaM [~~4,28~~27] and STEM [26~~],~~]) are required to leverage models and data at different spatial scales~~, including~~. These include social contact networks, local and global individual mobility patterns~~ of individuals as well as~~, location-specific ~~transmissibility~~control interventions, and epidemiological characteristics of the infectious disease in question~~ and control intervention data and models~~:

- The population model for ~~the GLEaM~~a global epidemic simulation ~~engine~~system can be based, for example, ~~is based~~on the Gridded Population of the World project by the Socio-Economic Data and Applications Center (SEDAC) [25~~] and~~], which has a resolution of 15 × 15 minutes of arc.
- Mobility models can include long-range air travel mobility data, 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 has become available for a number of countries including the US [11], and South-East 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 [4~~10,11,14~~13,18].

- Epidemic models allow the user to specify epidemiological parameters ~~for~~that are specific of the infectious disease (such as ~~reproductive number~~transmissibility and seasonality), initial outbreak conditions (e.g. 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 infection, the parameters of a typical model (the modified Susceptible-Latent-Infectious-Recovered model described in [27]) ~~includes~~include (a) the infection rate of contracting illness when an individual interacts with an infectious person; (b) infection rate scaling factors for asymptomatic infectors and treated infectors; (c) average length of the latency period (in which the individual is infected, but not infecting); (d) probability of symptomatic vs. asymptomatic infections; (e) change in the travelling behavior after the onset of symptoms; (f) average length of recovery; (g) percentage of infectious individuals that undergo pharmaceutical treatment; and (h) impact (e.g. on the length of the infectious period) of the treatment.

The output of ~~an epidemic~~a simulation is a multi-variate 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.

1.2 *Challenges*

While large-scale epidemic simulation systems such as GLEaM [~~4,28] and~~27] or STEM [26] ~~are~~represent very powerful and highly modular and flexible epidemic spread simulation ~~software~~systems, their power for real-time decision making could be enhanced by addressing the following challenges:

(a) *Complexity of the simulation and observation data*. A sufficiently useful disease spreading simulation system requires models, including social contact networks, local and global mobility patterns of

individuals, and epidemiological parameters for the infectious disease (e.g., infectious period). Epidemic simulations track 10s or 100s of inter-dependent parameters, spanning multiple layers and geo-spatial frames, affected by complex dynamic processes operating at different resolutions. Moreover, an ensemble of stochastic epidemic realizations may include 100s or 1000s of simulations, each with different parameters settings corresponding to slightly different, but plausible, scenarios [1,7]. As a consequence, running and interpreting simulation results (along with the real-world observations) to generate timely actionable results ~~are difficult~~pose challenges.

(b) *Dynamicity of the real-world observations.* A major challenge in using data- and model-driven computer simulations for ~~disease spreading and for~~ predicting geo-temporal evolution of epidemics for managing health emergencies, such as the 2014-15 Ebola epidemic in West Africa, is that the data, models, and the underlying model parameters dynamically evolve over time ~~requiring~~. 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: (a) revisions involve incorporating the real-world observations as well as updated probability surfaces into existing simulations to alter their outcomes; (b) 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.

~~Unfortunately, because~~In order 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 ~~systems to adequately~~

and systematically interpret observations, understand the processes driving epidemic outbreaks, and assess the robustness of conclusions driven from simulations to provide validated epidemic response strategies to public health authorities within a realistic time to have a significant impact on disease control.decision support systems. Therefore, tools that help (a) executing large-scale simulation ensembles under a large number of diverse hypotheses/scenarios, and (b) 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 spatio-temporal spread at different spatial scales, 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.

2 EpiDMS System Overview and Use Scenario

The key characteristics of data and models relevant to data-intensive simulations include the following: (a) voluminous, (b) multi-variate, (c) multi-resolution, (d) multi-layer, (e) geo-temporal, (f) inter-connected and inter-dependent, and (g) often incomplete/imprecise. Moreover, data and models dynamically evolve over time, due to preventivecontrol actions taken by individuals and public health interventions, requiring continuous adaptation and re-modeling.

The novel epiDMS software framework [1] 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. The services provided by epiDMS include (a) indexing and metadata and/or similarity-based search of large ensemble simulation data sets, including extraction of salient features from the inter-dependent parameters, spanning multiple layers and spatial-temporal frames, driven by complex dynamic processes operating at different resolutions [29]; and (b) data analysis, including identification of unknown dependencies across the input parameters and output variables spanning the different layers of the observation and simulation data [16,24].Unlike other dynamic modeling platforms

such as Berkeley Madonna [0], the services provided by epiDMS include

- storage and indexing of large ensemble simulation data sets and the corresponding models; and
- search and analysis of ensemble simulation data sets to enable ensemble-based decision support [15,23,28].

The target user group for epiDMS include 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.

2.1 *System Overview*

The *epidemic simulation data management system* (epiDMS [1]) for managing the data and models for data-driven real-time epidemic simulations consists of three major components (Figure 1):

- *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 ~~in parallel using STEM simulation engine,~~using an external simulation engine. Note that epiRun is not specific to any disease model or simulation engine and 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 (i.e., time series for each output variable) then become inputs for the epidemic data and model store (epiStore), described next.
- *Epidemic data and model store* (epiStore) ~~ingests,~~ stores, and indexes the relevant data and metadata sets. The data ~~sets~~ and models relevant for modeling large-scale epidemics include the following:
  - Network layers: An epidemic simulation requires one or more layers of networks, from local and global mobility patterns to social contact networks.
  - Disease models, describing the epidemiological parameters relevant to a simulation and

the parameter dependencies necessary in the computation of the disease spread.

- Simulation time series: For a given disease study, researchers and decision makers often perform multiple simulations, each corresponding to different sets of assumptions (disease parameters or models) or context (e.g. spatio-temporal context, outbreak conditions, interventions).
- Disease observations: These 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 ~~maintains~~captures simulation metadata (simulation model, parameter values, connectivity graphs) and simulation outputs (time series) and provides data analysis (such as clustering, classification, event extraction) to support decision ~~making~~ 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.

- *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.

2.2 *EpiDMS Use Scenario*

Let us consider a ~~group of public health officials at the CDC who are to develop~~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 (e.g, healthcare capacity) including the availability and effectiveness of the interventions, public health

experts execute a large number of simulations using the epiRun simulation ensemble creation engine~~,~~ ~~which relies on STEM~~ 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 ~~intervention triggers.~~when and what interventions are to be applied. Given these, epiRun schedules the execution of these simulations ~~on a parallel cluster and,~~. The simulation metadata and ~~stores~~results are then read and stored in epiStore. Intuitively, each simulation result corresponds to a "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.*

2.2.1 Scenario-based Querying and Exploration

A basic functionality of the epiDMS system 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 pre-executed simulations, based on SEIR ~~and SIR~~(susceptible-exposed-infectious-removed) and SIR (susceptible-infectious-removed) epidemic models, where the input transmission rate parameter was set between 0.3 and 0.6, the recovery rate parameter was set to 0.5, 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 (AZ), California (CA), and New Mexico (NM), for an ~~8 months long~~epidemic simulation that lasts 8 months (*Please see the online supplement for the details of this query as well as a detailed description of the query and visual exploration interface provided by epiDMS*).

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.

2.2.2 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 1) identify the relevant parameters that characterize transmission ~~characteristics~~patterns in near real time, 2) forecast epidemic spread as the epidemic evolves, 3) assess potential impact of intervention scenarios. This platform also allows the user to perform simulation refinements by narrowing down the parameter space of "possible worlds" based on 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.

3 Conclusions

In this paper, we describe and illustrate with an example a novel epidemic simulation data management system (EpiDMS [1]) that supports the generation, ~~searching, visualizing~~search, visualization, and ~~analyzing~~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.

Acknowledgements

We thank ~~to~~ the members of the EmitLab at ASU for their contributions to the epiDMS system. Please see the footnotes for the funding information and the conflict of interest statement.

Formatted: List Paragraph, Space Before: 2 pt, After: 3 pt, Line spacing: Double

Formatted: Font: Not Bold, Not Expanded by / Condensed by

4 References

1. Epidemic Simulation Data Management System (EpiDMS). Available at: https://hive.asu.edu:8443/MVTSDB/?p=epidemic. Accessed 10 May 2016.
2. Abubakar I, Gautret P, Brunette GW, Blumberg L, Johnson D, Poumerol G, et al. Global perspectives for prevention of infectious diseases associated with mass gatherings. Lancet Infect Dis. Jan;12(1):66-74.
3. Anderson RM, May RM. Infectious diseases of humans. Oxford: Oxford University Press; 1991.
4. ~~Balcan D, Hu H, Goncalves B, Bajardi P, Poletto C, Ramasco JJ, et al. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. BMC Med. 2009;7:45.~~
5. ~~4.~~ Barrett CL, Eubank SG, Smith JP. If smallpox strikes Portland. Scientific American. 2005 Mar;292(3):42-9.
6. ~~5.~~ Cauchemez S, Ferguson NM, Wachtel C, Tegnell A, Saour G, Duncan B, et al. Closure of schools during an influenza pandemic. Lancet Infect Dis. 2009 Aug;9(8):473-81.
7. ~~6.~~ Centers for Disease Control and Prevention. Interim pre-pandemic planning guidance: community strategy for pandemic influenza mitigation in the United States—early, targeted, layered use of nonpharmaceutical interventions. Atlanta (GA): The Centers for Disease Control and Prevention; 2007.
8. ~~7.~~ Chao DL, Halloran ME, Obenchain VJ, Longini IM, Jr. FluTE, a publicly available stochastic influenza epidemic simulation model. PLoS Comput Biol. Jan;6(1):e1000656.
9. ~~8.~~ 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 Sep 7;224(1):1-8.
10. ~~9.~~ 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 Jan;4(1):e13.
11. ~~10.~~ Flahault A, Vergu E, Boelle PY. Potential for a global dynamic of Influenza A (H1N1). BMC Infect Dis. 2009;9:129.

12. 11. Germann TC, Kadau K, Longini IM, Jr., Macken CA. Mitigation strategies for pandemic influenza in the United States. Proc Natl Acad Sci U S A. 2006 Apr 11;103(15):5935-40.

13. 12. Hollingsworth TD, Ferguson NM, Anderson RM. Will travel restrictions control the international spread of pandemic influenza? Nat Med. 2006 May;12(5):497-9.

14. 13. Kenah E, Chao DL, Matrajt L, Halloran ME, Longini IM, Jr. The global transmission and control of influenza. PLoS One. 2011;6(5):e19515.

15. 14. Khan K, Arino J, Hu W, Raposo P, Sears J, Calderon F, et al. Spread of a novel influenza A (H1N1) virus via global airline transportation. N Engl J Med. 2009 Jul 9;361(2):212-4.

16. 15. Liu S., Garg Y, Candan KS, Sapino ML, Chowell G. NOTES2: Networks-Of-Traces for Epidemic Spread Simulations. AAAI Workshop on Computational Sustainability, 2015.

17. 16. Longini IM, Jr., Nizam A, Xu S, Ungchusak K, Hanshaoworakul W, Cummings DA, et al. Containing pandemic influenza at the source. Science. 2005 Aug 12;309(5737):1083-7.

18. 17. Warwick J. McKibbin. The Swine Flu Outbreak and its Global Economic Impact. Brookings. May 4th 2009. Available at http://www.brookings.edu/research/interviews/2009/05/04-swine-flu-mckibbin. Accessed May 10th 2016.

19. 18. Merler S, Ajelli M. The role of population heterogeneity and human mobility in the spread of pandemic influenza. Proc Biol Sci. Feb 22;277(1681):557-65.

20. 19. 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 Sep;7(9):e1002205.

21. 20. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 2008 Mar 25;5(3):e74.

22. 21. 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(22):pii: 19227.

23. 22. 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 Jul-Aug;214(1-2):70-2.

~~24.~~23. Schifanella C, Candan KS, Sapino ML. Multiresolution Tensor Decompositions with Mode Hierarchies. ACM Trans. Knowl. Discov. Data (TKDD), 8(2), Article No. 10, June 2014.

~~25.~~24. Siu A and Wong Y C R. Economic Impact of SARS: The Case of Hong Kong. MIT Press. 2004, 3(1), p. 62-83.

~~26.~~25. Socioeconomic Data and Applications Center (SEDAC). Columbia University; Available at: http://sedac.ciesin.columbia.edu. Accessed 10 May 2016.

~~27.~~26. STEM. The spatiotemporal epidemiological modeler project. Available at http://www.eclipse.org/stem. Accessed 10 May 2016.

~~28.~~27. Van den Broeck W, Gioannini C, Goncalves 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.

~~29.~~28. Wang X, Candan KS, Sapino ML. Leveraging metadata for identifying local, robust multi-variate temporal (RMT) features. IEEE International Conference on Data Engineering (ICDE) 2014: 388-399.

~~30.~~29. Wu JT, Cowling BJ, Lau EH, Ip DK, Ho LM, Tsang T, et al. School closure and mitigation of pandemic (H1N1) 2009, Hong Kong. Emerg Infect Dis. 2010 Mar;16(3):538-41.

30. Berkeley Madonna - Modeling and Analysis of Dynamic Systems. http://www.berkeleymadonna.com/. Accessed 10 July 2016.

Footnotes Page

[i] Funding statement: This work is supported by NSF grants NSF # 1318788 and NSF # 1518939.

[ii] Conflict of interest statement: The conflict of interest (COI) disclosure forms are attached. Authors report no competing interests related to this manuscript.

EPIDMS: DATA MANAGEMENT AND ANALYTICS FOR DECISION MAKING FROM EPIDEMIC SPREAD SIMULATION ENSEMBLES

(ONLINE SUPPLEMENT)

Sicong Liu

Arizona State University

s.liu@asu.edu

Silvestro Poccia

University of Torino

silvestro.poccia@edu.unito.it

K. Selcuk Candan

Arizona State University

candan@asu.edu

Gerardo Chowell

Georgia State University

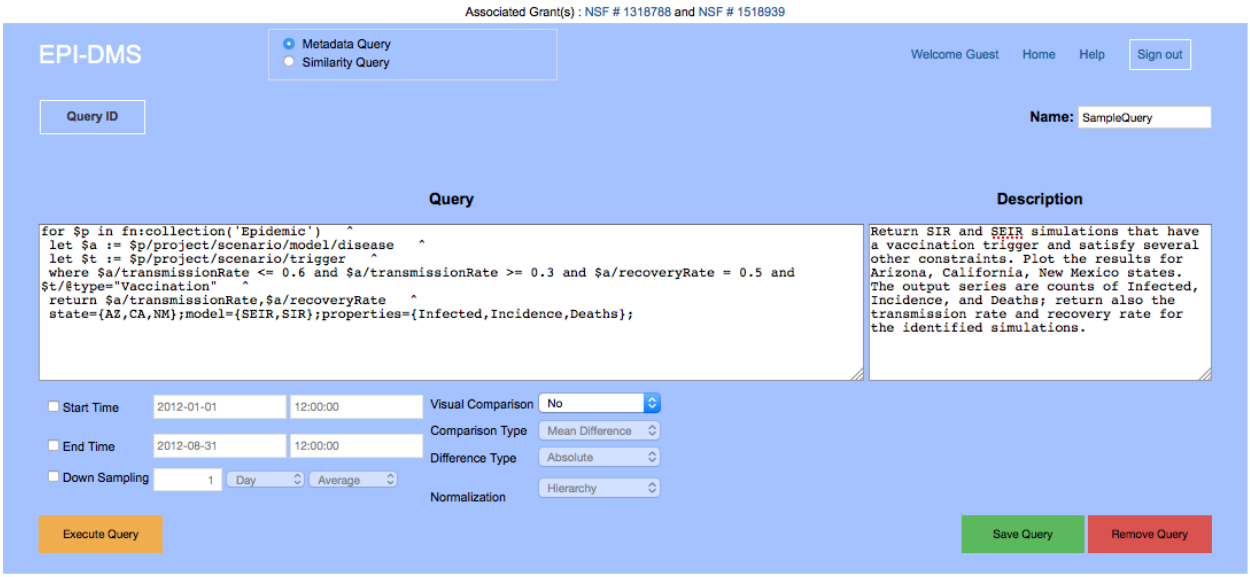
gchowell@gsu.edu

Maria Luisa Sapino

University of Torino

marialuisa.sapino@unito.it

A basic functionality of the epiDMS system is to retrieve epidemic simulations, stored in epiStore, based on the user specified scenario description.



The basic query interface, visualized above, provides the following functionalities:

- Query Menu --- visualizes the list of queries that are stored in the system.
- Query Box --- visualizes the selected query and/or allows the user to edit a query

- Query Description --- shows the description of the selected query and/or allows the user to add a query description.

EpiDMS provides a rich query language to specify user queries. Consider, for example, the following sample query:

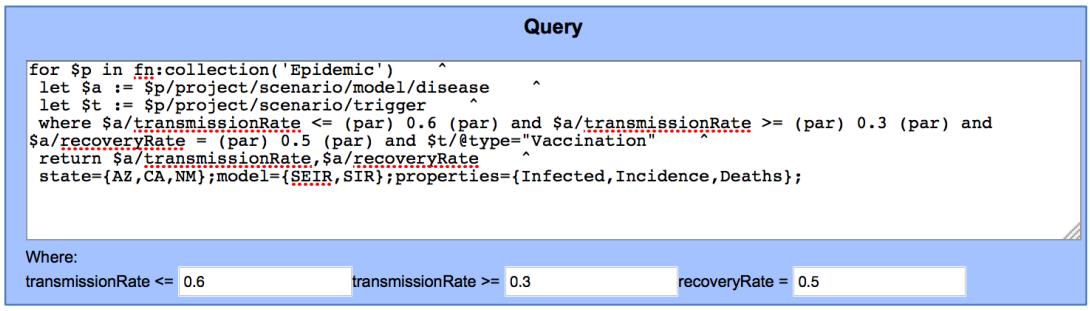
1. FOR $p in fn:collection('EpidemicSimulationEnsemble') ^
2. LET $diseaseModel := $p/project/scenario/model/disease ^
    - LET $triggerModel := $p/project/scenario/trigger ^
    - LET $epidemicScenario := $p/project/scenario ^
3. WHERE
    a. $diseaseModel/transmissionRate <= 0.6 and
    b. $diseaseModel/transmissionRate >= 0.3 and
    c. $diseaseModel/recoveryRate = 0.5 and
    d. $triggerModel/@type="Vaccination" and
    e. ($epidemicScenario/infector/@targetISOKey="US-~~NY~~CA" or $epidemicScenario/infector/@targetISOKey="US-NY") and
    f. ($epidemicScenario/graph = "mobility_graph_7.xml" or $epidemicScenario/graph = "mobility_graph_8.xml") ^
4. RETURN
    a. $diseaseModel/transmissionRate,
    b. $diseaseModel/recoveryRate,
    c. $epidemicScenario/graph ^
    d. STATE={AZ,CA,NM};
    e. MODEL={SEIR,SIR};
    f. PROPERTIES={Infected,Incidence,Deaths};
5. FROM ={01/01/2012 12:00:00}; TO={08/31/2012 12:00:00};
6. BY={1-D}; FUNCTION ={avg};

We describe the different components of this sample query below:

1. The “FOR” statement allows the user select the simulation dataset to query. In this example, the user selects to focus on the stored simulation set “EpidemicSimulationEnsemble”.
2. The “LET” statement allows to associate variables representing disease and intervention trigger models and epidemic scenarios.
3. The “WHERE” clause allows the user to specify conditions on the simulation models to filter those simulations that are relevant for the current analysis. In this example, the user specifies that for the returned simulations, the transmission rate parameter should be between 0.3 and 0.6, the recovery rate parameter should be set to 0.5, and that a “vaccination” type trigger should be included in the simulation model. The user also specifies that epidemic should have started at California (CA) or New York (NY) state and the “mobility_graph_7.xml” or “mobility_graph_8.xml” should have been used to generate the simulations.
4. The “RETURN” clause lists the simulation parameters to be returned in the result. In this example, the user is interested in the transmission rate, recovery rate, the mobility graph for each returned simulation. In addition, the query asks the system to return the time series corresponding to the “infected”, “incidence”, and “deaths” simulation output parameters for Arizona (AZ), California (CA), and New Mexico (NM) states.
5. In this clause, the user specifies that s/he is interested in only the first 8 months of the simulation.
6. Furthermore, the user specifies that the system return daily (1-D) averages of the simulation parameters for the specified duration.

1.1.1 Query Interface

The epiDMS query interface allows the user to specify and execute parametric queries. As illustrated below, parametric queries support query specification reuse – instead of writing a new query for different parameters, the user can specify and store a parametric query, which can then be invoked with different parameter values, as seen in the following example:

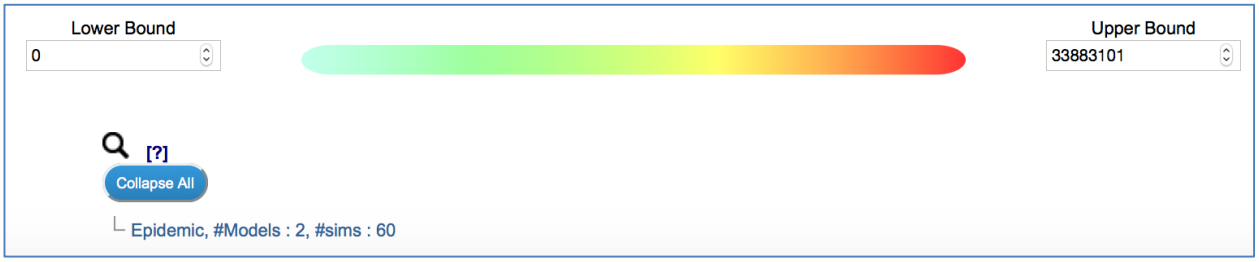


In the above example, those query parameters whose values are bracketed with the symbol "(par)" are interpreted as being parametric. The user can vary these values using a form-based interface without having to modify the source code directly.

1.1.2 Result Set Exploration Module

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 on this hierarchy and pick sets of simulations to study and compare in further detail the corresponding scenarios. This process is described below:

Once the matching simulations are identified, the user is presented with an initially collapsed hierarchy of results:

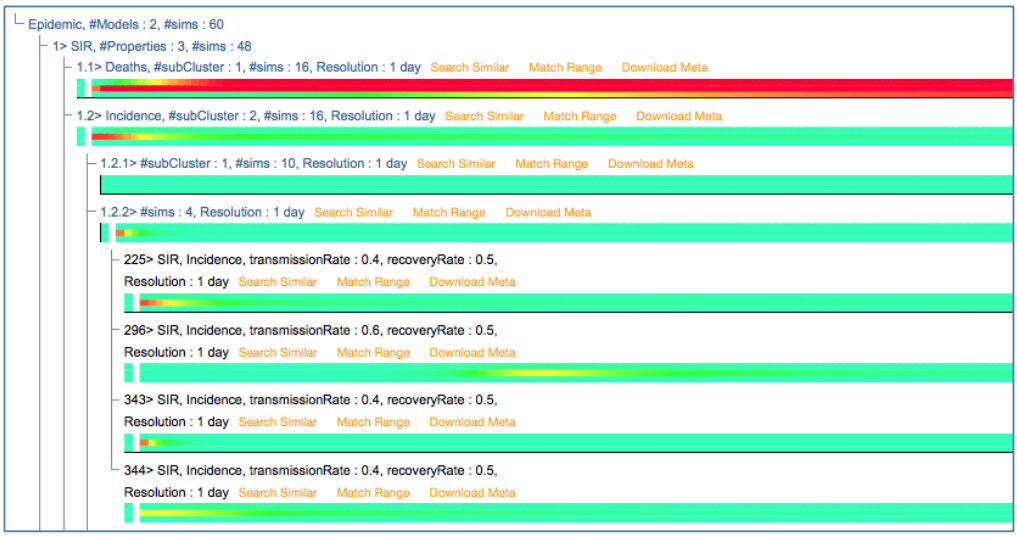


Above, we see that the query identified 60 matching simulations, from two different disease models. The legend at the top provides the scope of values in the results. The user can explore these simulations by drilling down or rolling up on the result hierarchy:

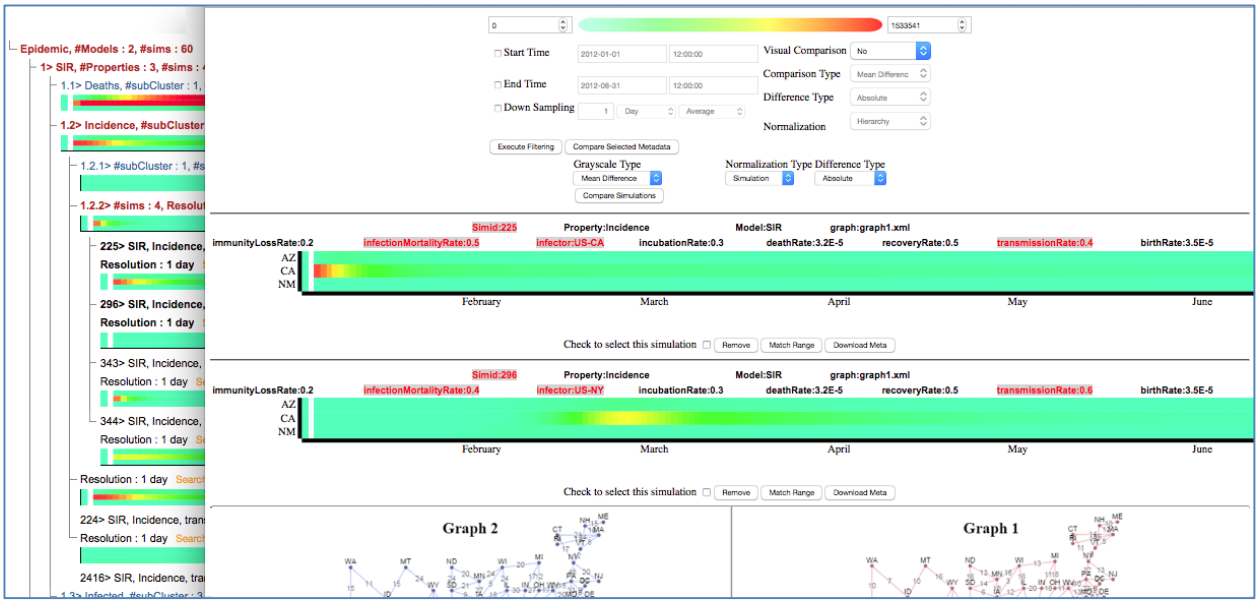
As we see above, the top-level of the result hierarchy includes the disease models (SIR and SEIR in this example). At the next level, the user is presented the output parameters specified in the query ("incidence", "deaths", and "infected" in this example). Under this level, the results are organized in the form of a cluster hierarchy, where similar simulations are clustered under the same navigation branch. For each node in the navigation hierarchy, a cluster representative is selected and the corresponding simulation is visualized in the form of a heatmap, where each row corresponds to a location (states "AZ". "CA", and "NM") in this example. The user can obtain detailed information about the presented simulations, by hovering the mouse on the heatmaps or download ~~metadata~~simulation results (in the form of CSV files) or metadata and model specifications (in XML format) corresponding to different simulations for further study~~.~~ or dissemination to decision makers.

The user can also use the "match range" feature to change the scale of visualization so that the upper bound of visualized values in the heatmap is modified in a way that matches a selected simulation to enable better visualization of its details. For example, in the example below, the heatmap scale has been modified to match the number of incidences, rather than the number of deaths; thus, we are able to better observe the differences among the incidence clusters:

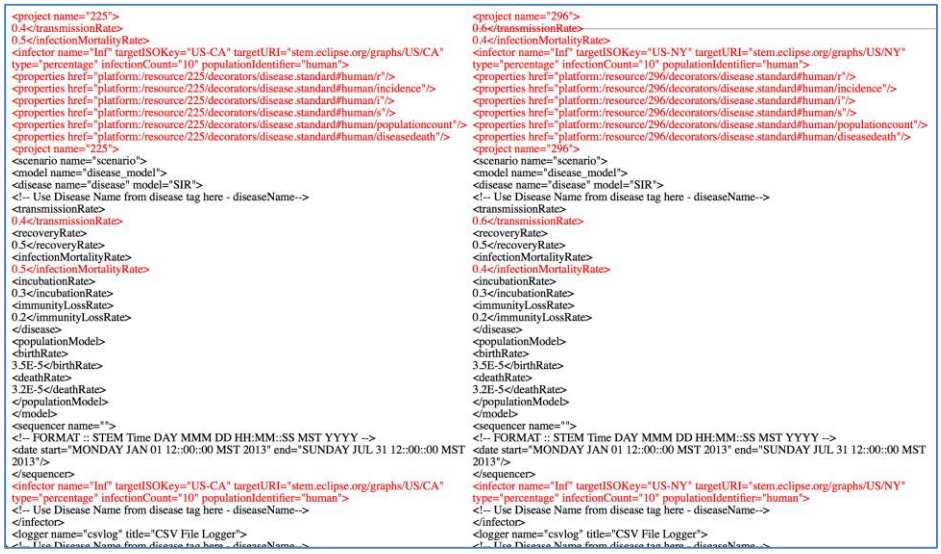
The user can explore these simulations by navigating on the hierarchy by drilling down or rolling up different branches. In the following example, the user has drilled down on the cluster 1.2.2 of the "incidence" data to observe the simulations clustered under this navigation node:



To further study individual simulations, the user then can double click on any simulation in the navigation hierarchy to place them in to a separate comparison interface. In the example, shown below, the user selected two simulations (#225 and #296) under cluster 1.2.2 to be studied and compared further in detail:



In the detailed comparison interface, the user can compare the two (or more) simulations side by side and observe the differences in the input parameters and models. The user can further ask the system to visualize the precise differences in the metadata corresponding to selected pairs of simulations:



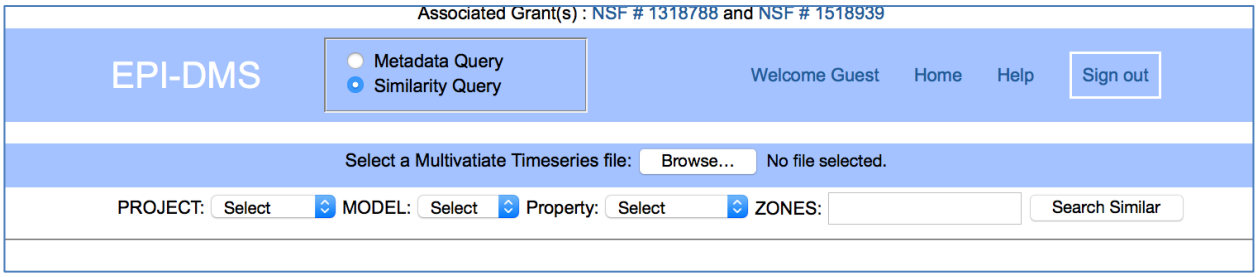
Here the text highlighted in red ~~point~~points to the differences in metadata corresponding to the pair of simulations selected or comparison.

1.1.3 Observational Similarity 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.

To use similarity based querying, the user can either click on the “Search Similar” option on the result visualization interface or switch to the “Similarity Query” interface and provide a file which contains the observations of interest:



Once a simulation/observation and states of interest are provided, the system searches in the databases existing simulations that show a similar pattern. Results are ranked in terms of their similarities to the provided query pattern:

Note that, once again, the user can obtain detailed information about the presented simulations by

hovering the mouse on the heatmaps or download simulation data or metadata corresponding to different simulations for further study. Moreover, as before, to further study individual simulations, the user can double click on any simulation in the navigation hierarchy to place them in to the comparison interface. Please see the accompanying video at https://www.youtube.com/watch?v=9w-4nDhXv3k for more details.

Frequently Asked Questions

***Question #1**: "It appears that epiDMS would be operated by those with at least moderate infectious disease modeling experience. Is it true that epiDMS requires programming skills by the operator (while there appears to be a GUI, there also appears to be a moderate amount of programming involved in operating this)."*

Answer: The target user group for epiDMS include a range of public health researchers and decision makers. While creation of models for ensemble simulations and formulating queries over ensembles simulations require moderate infectious disease modeling experience and familiarity with (not programming, but) declarative querying, epiDMS also provides parameterized queries and other interactive user interfaces to enable decision makers with minimal experience to explore large ensemble simulations.

***Question #2:** "Can you give a pathogen-specific example of a public health emergency in which the data, models and underlying model parameters dynamically evolve over time requiring continuous analyses and interpretations of the incoming data and adaptation of the networks and models."*

Response: The 2014-15 Ebola epidemic in West Africa was an example of such an health emergency where the situation (what we new about the disease characteristics, available and implemented intervention strategies, population dynamics, and social interactions among and within effected populations) continuously changed as the epidemic evolved, requiring reassessment and revisions models and re-interpretations of the data.

Question #3: *"How does epiDMS differ from existing modeling platforms and packages (e.g,. Berkeley Madonna or R)."*

Answer: Unlike other dynamic modeling platforms such as Berkeley Madonna, the services provided by epiDMS include

- storage and indexing of large ensemble simulation data sets and the corresponding and models; and
- search and analysis of ensemble simulation data sets to support ensemble-based decision support.

In that sense, epiDMS is less of a modeling tool and more of a multi-model, multi-instance ensemble simulation-based decision support system.

Question #4: *"Is epiDMS specific to a particular disease model or simulation engine? If not, how does different models fit within the database?"*

Response: We thank the reviewer for bringing to our attention that the original manuscript did not make it sufficiently clear that epiDMS is a model independent system by design:

- epiRun, for execution ensemble simulations, is not specific to any disease model or simulation engine and can wrap –as a black-box software component– any epidemic simulation engine as long as it provides command line invocation.
- epiStore, which stores epidemic models and the generated simulation ensembles, 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 is available. This wrapper based design ensures that models and simulations generated by different engines and tools can be imported into epiStore and queried and analyzed simultaneously irrespective of their origin.
- Finally, epiViz, which provides a web-based query and result visualization interface to support

user interaction and exploratory decision making is also model independent. More specifically, the underlying query specification language can support queries based on any model, without having to make any a priori assumptions regarding what the input and output parameters of the simulations are. Once they are imported into epiStore, parameters of any model can be queried, visualized, and explored.

The current alpha version of the system provides wrappers for the STEM simulation engine and can import models and simulations generated by STEM tool. The beta version of the tool will include wrappers for other systems.

Question #5: *"(i) What are the computational demands of epiDMS. e.g., can this be run on a standard laptop? A tablet/smartphone? From the video, it appears this is a web-based platform, but is there a stand alone downloadable form which can be run in potential areas with no internet connection (e.g., in certain field settings)?"*

Answer: The user interface of epiDMS is indeed a web-based platform and can run on any networked laptop and most tablets or smartphones. The backend, however, runs on server hardware. It is, however, possible to configure a laptop to act both as the backend and frontend.

Question #6: *"What is the speed of the simulation analyses?"*

Answer: This depends on the size of the simulation ensemble, number of variates/parameters of interest, the type of analysis, and the hardware configuration (memory, number of cores) at the back-end server platform. Having said that, we are doing our best to provide a near real-time and interactive experience to the users.

Question #7: *"What is the format of the modelling output? Can it easily be downloaded and disseminated to decision makers in public health practice?"*

Answer: Users of epiDMS can download simulation results (in the form of CSV files) or metadata and model specifications (in XML format) corresponding to different simulations for further study or

dissemination to decision makers.

Question #8: *"Can you confirm if this is a free system?...is there an open-source version of the software with scope for a community of developers?"*

Answer: An alpha version of the source-code for epiDMS is currently available upon request, and free of charge, to researchers and educators in the non-profit sector, including institutions of education, research, and government laboratories under an Apache 2.0 license (http://www.apache.org/licenses/LICENSE-2.0). The terms of the license allows individuals to modify the source code and to share modifications and also enable open source development of the software by other individuals and teams. The terms of software availability permits the commercialization of enhanced and customized versions of the software and incorporation of the software or pieces of it into other software packages. The beta release of the source-code will be available to the public through GitHub under the same terms.

Question #9: *"Is there a user-group forum for users to ask questions, trouble-shoot, show applications etc.?"*

Answer: While such a user-group forum does not currently exist, we will bootstrap a group along with the beta release of the system. In addition, we are planning to

- carry out demonstrations of epiDMS,
- give tutorials, and
- organize workshops

at leading forums targeting public healthcare researchers, scientists, and decision makers.