Published in final edited form as:

ACM BCB. 2015 September ; 2015: 156–165. doi:10.1145/2808719.2808735.

EpiCaster: An Integrated Web Application For Situation Assessment and Forecasting of Global Epidemics

Suruchi Deodhar,

Network Dynamics and Simulation Science Laboratory, VBI, Virginia Tech, Blacksburg, VA

Keith Bisset,

Network Dynamics and Simulation Science Laboratory, VBI, Virginia Tech, Blacksburg, VA

Jiangzhuo Chen,

Network Dynamics and Simulation Science Laboratory, VBI, Virginia Tech, Blacksburg, VA

Chris Barrett,

Network Dynamics and Simulation Science Laboratory, VBI, Virginia Tech, Blacksburg, VA

Mandy Wilson, and

Network Dynamics and Simulation Science Laboratory, VBI, Virginia Tech, Blacksburg, VA

Madhav Marathe

Network Dynamics and Simulation Science Laboratory, VBI, Virginia Tech, Blacksburg, VA

Suruchi Deodhar: suruchi@vbi.vt.edu; Keith Bisset: kbisset@vbi.vt.edu; Jiangzhuo Chen: chenj@vbi.vt.edu; Chris Barrett: cbarrett@vbi.vt.edu; Mandy Wilson: mandywil@vbi.vt.edu; Madhav Marathe: mmarathe@vbi.vt.edu

Abstract

Public health decision makers need access to high resolution situation assessment tools for understanding the extent of various epidemics in different regions of the world. In addition, they need insights into the future course of epidemics by way of forecasts. Such forecasts are essential for planning the allocation of limited resources and for implementing several policy-level and behavioral intervention strategies. The need for such forecasting systems became evident in the wake of the recent Ebola outbreak in West Africa.

We have developed EpiCaster, an integrated Web application for situation assessment and forecasting of various epidemics, such as Flu and Ebola, that are prevalent in different regions of the world. Using EpiCaster, users can assess the magnitude and severity of different epidemics at highly resolved spatio-temporal levels. EpiCaster provides time-varying heat maps and graphical

Additional authors: Bryan Lewis (Network Dynamics and Simulation Science Laboratory, VBI, Virginia Tech, Blacksburg, VA. email: blewis@vbi.vt.edu)

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. *BCB '15*, September 9–12, 2015, Atlanta, GA, USA.

Categories and Subject Descriptors

I.6.3 [Simulation and Modeling]: Applications; H.4 [Information Systems Applications]: Miscellaneous

General Terms

Design

plots to view trends in the disease dynamics. EpiCaster also allows users to visualize data gathered through surveillance mechanisms, such as Google Flu Trends (GFT) and the World Health Organization (WHO). The forecasts provided by EpiCaster are generated using different epidemiological models, and the users can select the models through the interface to filter the corresponding forecasts. EpiCaster also allows the users to study epidemic propagation in the presence of a number of intervention strategies specific to certain diseases.

Here we describe the modeling techniques, methodologies and computational infrastructure that EpiCaster relies on to support large-scale predictive analytics for situation assessment and forecasting of global epidemics.

1. INTRODUCTION

In recent times, there has been substantial interest in developing computer models to forecast various social and epidemiological events. This trend is driven by progress in data analytics and the availability of increasing amounts of data from various sources. Some examples of this include models to forecast social unrest events [23], models for predicting crowd behavior [13] and so on.

Several models have been developed for modeling and representation of Flu outbreaks in different regions of the world. Some notable examples include Google Flu Trends, developed by Larry Brilliant and his colleagues, that uses search terms entered by users, to develop statistical models for influenza forecasts [11]. Researchers like Butler et al. [8] have explained some of the shortcomings of GFT-based predictions by analyzing the results produced by Google Flu Trends. Research in Flu forecasting also includes work by Shaman et al. [27], who have developed real-time forecasting model of seasonal influenza outbreaks and applied it to derive retrospective ensemble forecasts using techniques similar to numerical weather predictions. Yang et al. [33] have described inference systems for influenza that can handle the observational biases and errors arising from the epidemic surveillance data gathered from the Centers for Disease Control and Prevention (CDC) and Google Flu Trends (GFT). They have used Bayesian Inference techniques to derive such forecasts. Osgood et al. [19] have demonstrated use of particle filtering techniques to robustly tolerate the noise in the epidemiological data feeds used in forecasting simulations, without making significant assumptions on the epidemiological processes used. Work by Vespignani et al. [31] includes using Monte Carlo Maximum Likelihood analysis for generating ensemble forecasts for different regions in the northern hemisphere.

There are also other models of flu forecasting using data sources like Twitter, developed by Sugumaran et al. [29], Tamerius et al. [30] and so on. Paul et al. [20] show the use of social media messages for predicting a variety of ailments and conditions, including obesity, allergies, and other ailments. These models have shown that health and epidemiological events can be forecasted using a combination of mathematical models and diverse data sources. There has been a flurry of activity in this field, including work by Yuan et al. [34], Nsoesie et al. [16], Chakraborty et al. [9], and so on. Recently, Paul et al. [21] have shown that use of Twitter actually improves the Flu forecast, reducing forecasting error by 17–30 percent over baseline models using only historical data. They have also demonstrated that

using Twitter data provides better forecasts compared to GFT data. Hickmann et al.[12] have demonstrated the use of Wikipedia access logs together with the weekly CDC reports to derive forecast results for Flu ahead of the peak flu season. There are several such Flu forecasting models developed by different groups. Nsoesie et al. [18] provide a comprehensive literature review on forecasting techniques used for studying the dynamics of influenza outbreaks. Yang et al.[32] have provided a summary of the comparison of forecast results obtained using six different filtering techniques for making retrospective Flu predictions including the maximum likelihood estimation via iterated filtering (MIF), ensemble filters such as the ensemble Kalman filter (EnKF), the ensemble adjustment Kalman filter (EAKF), and the rank histogram filter (RHF).

Research in epidemic forecasting models is not limited to Flu alone. Several models have also been developed for other types of infectious diseases such as Ebola and Malaria. The recent outbreak of Ebola in West Africa made headlines for its large number of casualties. Public policy makers needed tools and techniques to forecast the epidemic spread in the affected countries and other regions of the world. Epidemiological models were developed to assist decision makers for planning purposes. Some researchers including Lofgren et al. [14] and Rivers et al. [24] have argued that mathematical models are the key tools for Ebola epidemic response. Rivers et al. [25] modeled the effects of intervention strategies in containing the spread of Ebola in Sierra Leone and Liberia. Researchers including Merler et al. [15] studied the effectiveness of non-pharmaceutical interventions in controlling spatio-temporal propagation of Ebola. Researchers have also used dynamic modeling and Bayesian inference methods based on observations to generate weekly forecasts of the Ebola outbreaks in Guinea, Liberia and Sierra Leone. This includes the work by Shaman et al. [28]. Use of species distribution models for mapping the Ebola virus in Africa has been demonstrated by Pigott et al. [22]. Our lab, the Network Dynamics and Simulation Science Laboratory at Virginia Tech has been one of the leading research organizations in conducting extensive research in modeling the spread of Ebola in West Africa [3]. We have incorporated some of these models developed in our lab along with other models to study the Ebola outbreak along with other diseases within EpiCaster.

With the design and development of EpiCaster, our goal is to leverage such large and diverse sets of epidemiological models for generating epidemic forecasts for various diseases in different regions of the world, and make it available to the public health decision makers through a common, intuitive Web-based system. We have developed a platform that is not only able to handle models for current set of diseases, such as Flu and Ebola, but is also flexible enough to incorporate models for new diseases as they appear. The system is also extendable to add new global regions for which infection data becomes available. Through the user interface of EpiCaster, users can select the region, disease, models of computation, intervention strategies, and the source of surveillance data used in the forecast generation process. Our system supports collection of surveillance data through several different sources, such as the CDC, WHO, social media, and so on. Our modeling implementation allows use of causal models with detailed information, such as behavioral adaptations by individuals of a population and effects of policy-level interventions on disease propagation.

1.1 Our Contributions

EpiCaster is a public web application for situation assessment and forecasting of various infectious diseases in different regions of the world. The current version of the system is operational for Flu and Ebola. However, we have built EpiCaster in such a manner that it is extendible to other diseases prevalent in different regions, which can be incorporated into the system once epidemic surveillance data is available.

EpiCaster can be thought of as analogous to websites such as weather.com that provide users with near real-time weather forecasts. Using EpiCaster, users can view the current and forecasted state of various epidemics through high-resolution maps and plots at different spatio-temporal levels. EpiCaster supports the following classes of users: (i) casual users seeking general information on different epidemics such as Flu and Ebola, and (ii) epidemiologists and citizen scientists, who need to analyze the current and future states of epidemics to evaluate and test appropriate mitigation strategies.

There have been some systems published in the literature that produce epidemic nowcasts and forecasts over the Web such as the systems developed by Columbia University's Mailman School of Public Health[1] and HealthMap, developed by Brownstein et al.[2] at Harvard. These systems are primarily tied to a single model of forecasting of a given epidemic. It is not possible to analyze forecast results from these systems based on a range of models or surveillance sources, or to analyze the effects of mitigation strategies on epidemic propagation in future. This feature is particularly important for public health decision makers. The distinguishing feature of EpiCaster is that it is not tied to a single data source or forecasting model. EpiCaster can support a variety of forecasting models of epidemic propagation, including fine-grained models such as agent based models (ABM models). Hence, EpiCaster allows a detailed study of global epidemic forecasts for a range of diseases. In addition, EpiCaster also produces forecasts in the presence of targeted interventions strategies applied to specific subpopulations of a region, which may be selected by users. This provision allows advanced predictive analytics, which is not present in other existing systems described in the literature.

The design and development of the forecasting pipeline that generates high-resolution forecasts of various epidemics within EpiCaster, is a key contribution of our work. The class of epidemiological models used for forecasting are unique and make the system design challenging. We support not only the aggregate differential equation-based models, providing forecasts at a higher level of granularity, but also individual-based models that closely represent actual population and epidemic dynamics between individuals. The use of individual-based models allows fine-grained temporal, social and geographic resolution of forecasts.

Lastly, the overall architecture of EpiCaster, which provides flexibility to augment new data sources and diseases to the system, is a unique contribution of this work. We have developed flexible web services based on REST protocol to decouple the back end implementation from the user interface. We have designed and implemented several classes of RESTful web services that range from services to access atomic data items to services for gathering complex information such as epidemic counts for different diseases aggregated at multiple

spatio-temporal levels. The implementation of these flexible web services, along with optimizations to improve performance of the web services, is a significant part of the EpiCaster architecture.

1.2 Background

The conceptual design of EpiCaster has been developed over the course of several years. The first version of the application was designed specifically for supporting situation assessment and forecasting of Flu, and was called “FluCaster”. FluCaster was developed to support a variety of forecast models at the back end for computing and generating Flu forecast results. We leveraged some of the initial research work in the domain of epidemiological models for supporting Influenza, such as work by Shaman et al. [27], Nsoesie et al. [16], Tamerius et al. [30], and so on.

The initial version of FluCaster was designed using individual-based models of Flu forecasting. We implemented individual-based models for different states of the US. This version used synthetic population modeling, based on the research by Barrett et al.[4] by building synthetic social contact networks for all the states of USA. This data was used together with surveillance information obtained from the CDC for Flu forecasting. This version was successfully used for a number of experimental studies requested by CDC for forecasting of Flu parameters in different regions of the US, such as peak number of infections, peak time, and start and end of the Flu season.

Over time, new studies were requested by Intelligence Advanced Research Projects Activity (IARPA) and other organizations to conduct Flu forecasting for Latin American countries. Since detailed demographic and geographic information was unavailable for some of the Latin American countries, it was difficult to build synthetic information systems like that for USA and deploy individual-based models. However, we could develop other models such as compartmentalized models, also known as aggregate-based or Ordinary Differential Equation based models (ODE-based models) for Latin American countries to compute aggregate forecasts, which do not require fine-grained demographic information. Even though the model of execution was different, these models could provide forecast data output for visualization at an aggregate level. Hence, we decided to integrate and incorporate ODE-based models into the FluCaster back end so that users could select them while evaluating results for available regions in Latin America.

The main motivation to extend FluCaster to other diseases arose from the requirement last year to forecast the propagation of Ebola in West African nations during one of the worst epidemics in the world. Computational scientists developed detailed models of Ebola epidemic propagation based on disease dynamics, such as infection through bodily fluids, possible infections through deceased individuals, and so on [26]. Our lab produced forecast results every week based on these models and provided them to IARPA.

From this experience of producing forecast results for Ebola, we realized that the forecasting pipeline of Ebola would work similar to that of Flu, from a data representation and systems standpoint, producing some aggregate form of forecast output. The modeling of the disease dynamics and the network of Ebola is vastly different from that of Flu, since the propagation

dynamics are different. However, the forecasts of Ebola could be made available to end users done in a similar way. This led to the conceptualization and design of “EpiCaster”, for supporting a range of global epidemics.

2. CONCEPTUAL ARCHITECTURE

The architecture of EpiCaster has been designed to support scalability with respect to large volumes of data, gathered from multiple surveillance sources, and manipulated using epidemiological models to arrive at forecast results. These forecasts can be viewed at multiple geographic, social and temporal resolutions.

Figure 1 shows the conceptual architecture of EpiCaster. The backbone of EpiCaster’s architecture can be described in four parts. First, information is gathered using various surveillance instruments, such as the CDC, local hospitals, social media sources and so on. For a disease like Ebola, where surveillance information from affected countries in West Africa is very scarce, non-traditional surveillance methods can be used and incorporated into the system. For instance, for Ebola, we use information collected from WHO and Ministries of Health (MoH) of affected countries, about current infection trends. EpiCaster builds on the premise that this collected information provides a good representation of the extent of the current epidemic. We use this information as input to the forecasting pipeline to produce forecast results. In addition, EpiCaster uses synthetic contact networks as the basis of population demographics of different regions and as a fabric of disease transmission, especially for individual-based epidemic models.

The second part of the conceptual architecture involves forecast modeling and generation. We use complex epidemiological models, including individual-based models and aggregate models, as implemented using multi-agent simulations as the key components of the prediction pipeline.

The third component of the EpiCaster architecture involves storing the forecast output in a central data repository so that it can be retrieved in an effective way. The data store needs to handle the flexibility of querying the database to retrieve forecasts based on any combination of regions, diseases, interventions, models and data sources, while still supporting scalability. Hence, the data representation schema is critical.

Lastly, the forecast data retrieved from the data store needs to be passed back to the user interface (UI) so that it can be rendered on the user’s web browser. The middleware that we have developed is able to handle flexible inputs from the users, including disease, region, models, data sources, and interventions. It also handles large scale data transfers to visualize current epidemic and forecast data on heat maps and plots.

3. EPIDEMIC SURVEILLANCE AND POPULATION DEMOGRAPHICS

One of the most important components of the epidemic forecasting process is collection of up-to-date information on the extent of current epidemics. This is termed as epidemic surveillance. In addition to current epidemic trends, it is also important to have information on population demographics of different regions including age-groups and household

income, that determines the extent of control strategies that may be applied for containing an epidemic. Hence, the data needed as input for epidemic forecasting in EpiCaster primarily falls under the following categories.

- Population demographic and contact data: In EpiCaster, we use demographic data corresponding to populations in different regions of the world that is derived from the synthetic contact network generation process. This synthetic population generation process is described in detail in [4, 5, 10]. This data generation process takes into account mobility of individuals based on their activity patterns in different regions.
- Surveillance data: Surveillance data within EpiCaster includes spatial and temporal epidemic time-series data. In EpiCaster, surveillance data is gathered from different sources depending on the disease and the region in which it is prevalent. For instance, our implementation uses surveillance data from CDC for Flu. We also use GFT as an epidemic surveillance source, even though GFT data itself is an estimate of Flu surveillance data based on a model built using Google search terms. For Ebola, the surveillance information is scarce and is gathered from the Ministries of Health (MoH) and the WHO.

The synthetic population and contact network data are more or less static in nature, except for changes related to interventions. As an epidemic evolves, more individuals may get infected and some of them may recover from the disease. Hence, the surveillance data related to epidemics is dynamic in nature. This is reflected in the dynamic nature of the data obtained through sources such as the CDC.

EpiCaster collects all the surveillance related data on multiple epidemics together with the demographic data in a common data store. This data comes in different formats such as text files, csv files, json format and so on. We have developed a data enrichment module that processes the collected surveillance data in different formats and transforms it, so that it can be accessed by the prediction algorithms. This data is then used by different forecasting models that are part of the EpiCaster backend to generate forecasts of epidemics in different regions.

4. FORECAST MODELING AND GENERATION

Forecast modeling and generation is the most important component of the EpiCaster architecture. We have developed and designed the backend infrastructure of EpiCaster that implements the forecast modeling and generation processes.

The back-end infrastructure of EpiCaster is responsible for generating forecasts for different epidemics based on high-resolution epidemiological models. The back end handles execution of large-scale simulations to generate and store epidemic forecasts.

The main components of the back-end infrastructure of EpiCaster include:

1. An epidemic forecasting pipeline that executes multi-agent epidemic simulations based on surveillance data, and uses a non-linear optimizer for forecast matching and generation.
2. A digital library that is a complete repository of information related to different current epidemics.

We describe the epidemic forecasting pipeline in the following section and describe the digital library that stores the epidemic forecast results in the Forecast Storage and Retrieval Section 5.

4.1 Epidemic Forecasting Pipeline

The EpiCaster epidemic forecasting pipeline is built to support a variety of epidemiological models and techniques that can derive epidemic forecasts based on advanced stochastic computations and optimizations. We describe these computations and optimizations in the following sections.

The epidemiological models and simulations that generate forecasts are the main building blocks of the pipeline. The pipeline is developed in such a way that it provides several techniques and methods for translations and transformations to handle the discrepancies in data generated from various models.

We use a simulation-driven optimization approach for the forecasting pipeline. From a modeling standpoint, this method has been described in detail by Nsoesie et al[17]. Our implementation of this approach has three main components:

- i. high resolution models of infectious diseases and their corresponding simulation engines. The models may be individual-based simulations or aggregate-based simulations.
- ii. a stochastic programming-based method for parameter matching and searching based on input epidemic surveillance data.
- iii. a model library comprised of a large number of plausible precomputed outcomes of epidemic forecasts using the first two steps.

4.1.1 Epidemic Models—EpiCaster supports both aggregate-based models, implemented using ordinary differential equations (ODE); and agent based models (ABM) which represent disease propagation in a population through person to person interactions. The aggregate-based models are usually called compartmental models. Both these models produce stochastic forecasting outcome, but the granularity of data involved in the computation and the forecast results is different.

In an ODE-based model, the population (N) is partitioned into different compartments. The SEIR model is one type of an ODE-based model where the population is divided into four main compartments – Susceptible (S), Exposed (E), Infectious (I), and Removed (R). Within each compartment, people are homogeneous and fully mixed. If β is the transmission rate, α is the rate at which an exposed person becomes infectious, and ν is the removal rate, then the following equations describe the dynamics of the epidemic in the population. An ODE-

based simulation based on these equations computes number of Susceptible (S), Exposed (E), Infectious (I), and Removed (R) individuals at each time step (typically represented as a day). So, the model is able to produce aggregated counts of infected individuals in different regions every day. The model is also able to assimilate forecasting in the presence of interventions applied on a subset of the population, given by a certain percentage of the population.

$$\frac{dS}{dt} = -\beta \frac{I}{N} S \quad (1)$$

$$\frac{dE}{dt} = \beta \frac{I}{N} S - \alpha E \quad (2)$$

$$\frac{dI}{dt} = \alpha E - \nu I \quad (3)$$

$$\frac{dR}{dt} = \nu I \quad (4)$$

$$S + E + I + R = N \quad (5)$$

In an ABM-based model approach, the population is represented by a network where the nodes are represented as individuals and the edges are the contacts between the individuals. The nodes are heterogeneous and have demographic, geographic, and economic attributes. The edges represent contacts due to physical proximity and have durations associated with them. The disease transmits from node u to v only if u is infectious, v is susceptible, and u has a contact with v . The probability of transmission is

$$p(u, v) = 1 - (1 - \tau)^{d(u, v)}$$

where τ is the probability of transmission in every unit time of contact, and $d(u, v)$ is the contact duration. An ABM simulation computes the disease transmission from node to node based on this probability.

There are multiple advantages of using the ABM-based model over the ODE-based model. Firstly, the ABM approach is able to represent fine-grained population characteristics of people within a region without assuming homogeneous mixing of population like the ODE-based approach. Secondly, ABM simulations allows representation of interventions for subset of populations based on demographic characteristics and not just a percent of the

population. Since each individual is represented as a node in the network, the pharmaceutical interventions such as vaccination decrease the probability of transmission on the nodes being intervened, while non-pharmaceutical interventions such as school closures remove the contact edges between the nodes being intervened. This allows representation of targeted interventions such as vaccination of adults in a certain zipcode, and school-closures in a particular block group of a county. Representing and analyzing such fine-grained interventions is not possible in the ODE-based approach. Thirdly, the type of metrics that can be evaluated using the ABM-based approach are much detailed than the ODE-based approach. For instance, the network-based models not only support forecast metrics such as aggregated infection counts, but also infection counts by demographic characteristics such as age-groups, gender and so on. This allows analysis of forecast results such as projecting the infections among females in a population, or forecasting the effects of an ongoing epidemic on school-age children.

The difficulty associated with the ABM simulation, however, is that it requires a realistic representation of the contact network for every region. Also, executing an ABM simulation is much more time consuming than its ODE counterpart because of the scale of data and computations involved. We have constructed synthetic populations and synthetic contact networks for several countries in the world, including the USA, Mexico, Liberia, Sierra Leone, and Israel. These provide a realistic representation of the actual global populations. We execute ABM simulation based on these synthetic populations. For countries where we have not built synthetic contact networks due to lack of information availability, we rely on the ODE-based simulations to compute epidemic curves from models.

4.1.2 Model Library—The simulation optimization process is a massive large-scale computation process. Hence, it is not just time-consuming, but also challenging to produce forecasts using this process in near-real time when surveillance data is available. To make the process more efficient, we have developed a precomputed model library that produces a large set of possible forecast outcomes considering a range of epidemic parameters. This pre-computation is done using Indemics[7], a data intensive, interactive, high performance modeling environment developed for pandemic planning which is an extension of EpiFast[6]. When new surveillance data comes in, this model library is searched, and the best match is used for further computations as described in the following subsection.

4.1.3 Parameter Matching and Search—We use a Bayesian approach for deriving the results of situation assessment and forecasting in different regions. Our model implementation is based on the simulation optimization approach to epidemic forecasting, described in [17].

With a given time-series obtained from surveillance on infection counts periodically (typically on a weekly basis), the pipeline either searches the precomputed epidemic curves in the model library (called the DL approach), or calls a derivative free optimization algorithm to search the model space; the purpose is to find a model that generates epidemic curves that are the most *similar* to the given surveillance curve.

Suppose $\mathcal{M} = \{(p_1, p_2, \dots, p_k) \in \mathbb{R}^k\}$ is the set of models, where p_i represents either disease parameters such as transmission rate and removal rate in the classical SEIR model, or intervention parameters such as compliance rate, efficacy of pharmaceutical interventions, duration of non-pharmaceutical interventions, timing of the interventions and so on. If $S \in \mathcal{S}$ represents an epidemic simulation model, which takes a model $M \in \mathcal{M}$ as input, then the output epidemic curve produced by the simulation is given by $S(M) = x = (x_1, x_2, \dots, x_T)$. This represents the number of infections in each week.

The epidemic simulation is a stochastic process and hence we have multiple epidemic curves from the simulation output. $S(M)$, given by $x_1, \dots, x_R$, represents infection counts at different geographical levels, per week across several replicates. Currently, we use avergaing techniques by finding a mean of the infection counts derived from several replicates and store the forecast output as an aggregated count of infections at different geographical levels within the specified region for which forecast is sought.

Let $y = (y_1, y_2, \dots, y_T)$ be the observed epidemic curve from surveillance. With given y , which has weekly infection counts, the DL approach and the optimization approach both try to evaluate models, either in the library or chosen by the optimization algorithm, by computing a distance measure between y and x . We have implemented Euclidean distance as a simple distance measure calculation between y and x in our current version, but the system supports a variety of distance measures to be used for this purpose. If a match is found, then the output infection time-series curve is used as the forecast outcome. If there is no match, then the epidemic parameters such as transmission rate and removal rate in the classical SEIR model are used from the closest matching model for running a simulation to derive the forecast results.

The epidemic forecasting pipeline of Epicaster consists of three main modules – the epidemic propagation engine, the comparison module, and the non-linear optimizer module. The epidemic propagation engine executes epidemic simulations based on a range of input epidemic time-series data, and stores the forecast outcomes in the model library along with their model parameters, including disease transmissibility, incubation period and infectious period. Since we use synthetic population data for ABM based models, we have detailed demographic information available for each region. With the granularity of demography information available, the forecast epidemic simulations can be executed with not just the base case of 'no interventions', but also with fine-grained interventions applied to certain sub-populations of the region. Some of the intervention strategies that can be supported for epidemic evaluation and forecasting include, vaccination of a sub set of individuals of a population and social-distancing of certain subset of infected individuals. For instance, the pipeline is able to compute infection counts in a particular block group of a city, when say 20 percent of school-age population is vaccinated, or when school-closure is enforced in a particular county of a region. The corresponding results of the epidemic simulation are stored in the model library and used for parameter matching when intervention information is available from surveillance. Support for such forecasting of epidemics, in the presence of targeted mitigation strategies, is a distinguishing feature of our system.

5. FORECAST STORAGE AND RETRIEVAL

EpiCaster allows forecasting of epidemics on certain demographic sub-groups of a population as well as in the presence of interventions applied to subpopulations of a region. To support such features, it is essential to store the forecast data in such a way that it can be retrieved easily, in the presence of different analytical constraints. In addition, the database storing the forecast results needs to support flexibility in terms of the range of analytical queries that it can handle. The typical type of data that needs to be retrieved using queries includes finding aggregated counts at regional, national and world levels per disease, and ranking of infection counts by region per disease. For instance, sample queries might include “Find the overall infection counts of Ebola in the world this week”, “Find the top 3 regions in USA that are forecast to have highest counts of population affected by Flu”, and “Find the Flu infection counts in Montgomery County, Virginia in the next week”.

We have implemented a relational database schema for storing forecast results, considering the flexibility and adaptability of analytical queries required in EpiCaster. The relational database schema, with its well-defined relational algebra allows a range of query plan execution. We store forecast results for each region at multiple geographical levels together with the demographic and intervention information related to each region. This allows querying the subset of forecast results from the forecast output in a flexible way.

We have also applied several mechanisms to the database schema design for improving the speed of query execution to retrieve forecast results faster. Some of the optimizations that we have implemented to streamline the epidemic data collection, forecast data generation, and retrieval process include denormalization, summarization, transformations, look-up registries and automations. We have precomputed the often-requested data and denormalized the data in the database wherever possible, for faster response times. With precomputations and denormalization, the response times have dropped from minutes to seconds. For instance, we have created a summary table containing precomputed data for the retrieval of aggregate-level totals derived from individual ABM-based models. We have developed a look-up registry that is an extension of the star schema design from the data warehousing literature, that allows looking up pre-computed forecast output based on different dimensions such as region, disease, intervention, surveillance source and forecasting model. Also, to avoid any manual interventions in the data collection and manipulation processes, we have automated the creation of forecasts and the generation of table summaries when surveillance data is gathered.

5.1 Data Flow and Scale of Data Within EpiCaster

Figure 2 shows the flow of data within EpiCaster from surveillance to forecast representation. As seen in the Figure, information about current infection trends is gathered from multiple surveillance sources such as the CDC, WHO etc. based on the disease type and epidemic information availability. This information is stored in various formats such as csv or text files. We have a module that converts this data into relational format and stores into the database.

The comparison module from the forecast pipeline, as explained before, matches the time-series curve from surveillance with the precomputed model library. The pipeline is executed based on whether the match is found or not found and the forecast results are stored in the database. The synthetic population information is also stored in the same database. We use a bunch of working tables for data manipulations and transformations.

The final forecast outcomes are stored in aggregate format in forecast summary tables, from which the forecast can be retrieved, for each region, for each disease, with or without interventions.

The surveillance data is obtained periodically (typically on a weekly basis) and the forecast needs to be computed as soon as the surveillance is available. With the pre-computations, aggregations and transformations from different formats, the scale of data produced is very large. For instance, in our current implementation of EpiCaster, we compute Flu forecasts for USA and India, and Ebola forecasts for Liberia and Sierra Leone. The scale of demographic and contact network data for these regions is in the range of around 250 GB. We obtain surveillance data every week for Flu and Ebola in different formats. Considering intermediate data involved in transformations and manipulations, the scale of this surveillance data is around 10 GB per week. We have ABM and ODE based simulations that produce forecast outcomes. This output forecast data is in the range of 15–20 GB. Overall, the current EpiCaster implementation stores around 500 GB of data, with around 20–30 GB of data getting added every week for Flu and Ebola. With addition, of new regions, models and diseases, there will be a corresponding increase in the scale of this data.

6. FORECAST AVAILABILITY OVER THE WEB

The forecasting output produced by the EpiCaster forecasting pipeline and stored in the database needs to be published in order to be consumed and displayed to the end users through the EpiCaster User Interface. The data stored in the back end cannot be tightly coupled with the UI applications because any change to the pipeline or the forecast generation process would affect the user interactions. To avoid such tight-coupling, it is essential to develop a middleware that can handle the interactions between the back-end infrastructure and the EpiCaster UI.

6.1 EpiCaster Middleware

The main requirements of the middleware for EpiCaster include

- the ability to provide access to epidemic data for flexible input types, including region, disease, intervention, model and source.
- the flexibility to be agnostic to back-end data changes and execution, either on HPC resources or pre-computed execution
- the ability to add/modify/remove back-end data without affecting service guarantees or requiring platform reengineering.
- the ability to be accessible by any consumer system, to be displayed on the Web or on social platforms.

To support these requirements, we have developed a Web services-based middleware architecture for EpiCaster that supports communication and interaction between its component systems, mainly the Web-enabled User Interface and the Back-end Infrastructure. We use a RESTful implementation of web services. The web services-based architecture is implemented using the Model View Controller (MVC) framework, where the controller accesses appropriate models, represented using DAOs (Data Access Objects). Data is passed back and forth from the web UI to the back end using JSON (Java Script Object Notation) format, which is a lightweight data interchange format. This design decouples the front-end user interface from the back-end infrastructure, and enables more flexibility and adaptability to EpiCaster. In this way, the Web UI can be modified without affecting the back-end system logic. Similarly, the API implementation may be modified at anytime without affecting the UI, as long as the JSON output published for communication remains the same.

The REST-based APIs provided by EpiCaster fall under the categories of Push API, Pull API and Query API respectively. For instance, the API, *getDiseases()* that returns all the diseases for which a forecast is currently available, is an example of a Query API, whereas *addQuestionnaireData(question, answer)* is an example of a Pull API. The Push API, *getTop3InfRegions(region, disease, intervention, week)* pushes back results to the requesting application and returns the top 3 infected regions in the input region, affected by a disease in the given week, when the intervention type passed as input has been applied on the subpopulation.

The API-based web services provided above hide the complexity of the back-end infrastructure of EpiCaster, and provide easy access to its services.

The performance of EpiCaster and its ability to support multiple simultaneous users accessing the system via the Web, is dependent on the performance of the backend infrastructure including the forecasting pipeline and the middleware that handles the flow of data within the system. In the following sub-section, we present the results of performance evaluation of the EpiCaster Web services. The performance evaluation results provide insights into the scalability of the application with growing epidemiological data.

6.1.1 Web-Service performance evaluation—The main criteria for evaluating the performance of web services include Quality Of Service (QoS) parameters such as response time, availability, throughput, and scalability.

Currently, the EpiCaster web services have been deployed on a virtual machine with a 2.93 GHz Intel Xeon processor. The virtual machine has 1 GB RAM and 122 MB reserved for cache space. The VM has 120 GB of disk space allocated to it.

Out of the total main memory available on the VM, the EpiCaster web server hosted on Tomcat uses 256 MB of RAM. The caching on the web server is handled through the Java internal caching mechanism.

With this configuration of web services, we expect the current deployment of EpiCaster to support about a 100 users using the system simultaneously (ramp-up period = 10 seconds).

We used Apache JMeter, an open source tool for conducting load tests, to evaluate the performance of the web services.

Experimental Setup

We conducted two sets of experiments. 1. To find out the number of users that the system can support simultaneously through the current deployment of EpiCaster without degradation, and 2. To find the effects of the size of the region on the performance of epidemic forecast output through the APIs.

For the first set of experiments, we used a Push API called *getOverallInfection()*, which retrieves the overall infections in different regions of the world based on the input disease type. This API is used for fetching information on the map after the user selects the disease to be studied. Table 1 summarizes the results of this experiment. The number of users accessing this API is kept constant at 100, but the ramp-up time is varied from 1 to 20 seconds. The results show the response times when 100 users access the API over the specified ramp-up time. As the ramp-up time is progressively reduced, the performance of the API decreases and the response time degrades. As can be seen in the Table 1, the current deployment of EpiCaster can support 100 users accessing the system within 5–10 seconds with no visible lags.

For the second set of experiments, we used 2 APIs – *getInfectionPlot()* and *getTop3InfectedRegions()*, which are used extensively in EpiCaster to fetch the plot of the epidemic forecast, and for getting the top 3 infected sub-regions in the given region respectively. We conducted experiments on the regions CA (Population: 33.6 million), VA (Population: 7.25 million) and WY (Population: 0.56 million) in USA; and on Delhi, India (Population: 13.9 million) for Flu. We also studied the APIs for Ebola for the Liberia region (Population: 4.2 million). For this experiment set, we used 100 users over the period of 10 seconds for each API test.

Table 2 summarizes the results of our experiments. From the experiment results, it is evident that the API response time and throughput of the API is independent of the size of the input region and the disease for which the forecast is sought. This shows the stability of our application to support the addition of any new region or disease without degradation in performance of the system.

6.2 EpiCaster User Interface

EpiCaster's Web-enabled User Interface allows users to select the disease for which the user seeks outbreak information; in the current version of the application, users can choose either Flu or Ebola, but this is extendable. Once the disease is selected, EpiCaster displays all the regions where the disease outbreak is prevalent on a map of the world. The user can then select a particular region on the map to get a fine-grained assessment of disease activity as well as forecasts within that region. Figure 3 shows a snapshot of EpiCaster displaying infection counts for Ebola in Bong county of Liberia.

EpiCaster's Web-enabled User Interface consists of the following main components.

- **Interactive Timeline:** EpiCaster provides a timeline covering a range of weeks over which a user may view epidemic information. The available timeline options include the current week and any of the last four weeks for situation assessment; users can also select the timeline for up to two weeks in the future to view forecasts.
- **Interactive heat map:** EpiCaster provides an epidemic heat map that shows infection levels aggregated at multiple geographical levels, such as state, county and blockgroup. GIS data is used to provide interactive maps to improve data visualization. Once the timeline is selected, EpiCaster displays epidemic activity in the selected region. The interactive heat map is displayed with different color schemes to allow the user to analyze the severity of the epidemic in the region. The user can also drill down to each subregion of the map to view infections at a granular level. The varying number of geographical levels and naming conventions in different regions of the world makes this visualization complex. To retrieve and display the forecasting results of each block group, in each county, in each state of a nation is a challenging task.
- **Plots (Epicurves):** EpiCaster allows the users to view Epicurves showing infection trends in a region over a time period, as well as the peak infection count in the specified time period. Users can select the plots to be viewed for the past and current weeks based on surveillance sources such as GFT and CDC, and view forecast plots derived from a range of epidemic models for the future weeks.

7. EXPERIMENTAL STUDY USING EPICASTER

We have used EpiCaster for performing several experimental case studies to evaluate the extent of ongoing epidemics in different regions of the world as well as to forecast future propagation. One of the most notable experimental studies for which we have used EpiCaster is the forecasting of the recent Ebola outbreak in West Africa. We used the ABM-based model simulation for situation assessment and forecasting of this epidemic in Liberia and Sierra Leone. The output Similar experiments have also been conducted using other ODE-based models and ABM-based models for Flu and Ebola.

In the 2014–2015 Ebola pandemic, it was observed that in mid-October 2014 and mid-December 2014 there were significant changes on the transmissibility of the disease. Given the flexibility of our modeling design in EpiCaster, we were able to incorporate this epidemic parameter into the system design and re-calibrate our model to study the effects of the pandemic propagation, taking into account this change.

We ran the EpiCaster forecasting pipeline for Sierra Leone in West Africa with and without the changes in transmissibility parameter and gathered the forecast results for analysis. The forecasting simulation model that we used is designed to produce forecast results up to six months in the future. We present our forecast results as follows in three plots: (i) in Figure 4 we made the prediction without considering either of the two transitions, so the transition

parameters are not considered in the model searching; (ii) Figure 5 is the prediction made with the only mid-October transition being considered in the model searching; (iii) and Figure 6 is the prediction made with both transitions being taken into the account in the forecasting simulation.

As can be seen from the Figure 4, when the transition parameters are not considered the model tends to over-estimate the number of infections in the forecast period. The number of infections computed by the model tend to be much higher than the observed number of infections. Figure 5 shows that the model slightly underestimates the number of infections in the forecast period, when a single transition is considered in the simulation. The curve in Figure 6 shows that when the model takes into account both transitions in the forecasting simulation, the observed infection count and model infection count are very close. This shows that when the most updated information is available from the environment, the model outcome has to be re-calibrated and the forecast outcome from EpiCaster can be improved to mirror the actual observed infections much ahead of time.

8. CONCLUSION AND FUTURE WORK

We have developed EpiCaster, an integrated system for situation assessment and forecasting of several global epidemics. Currently, the system is operational for Flu and Ebola. We plan to extend EpiCaster to other diseases such as Malaria, which is a vector-borne disease. New diseases can be easily incorporated into EpiCaster by making the forecast input and output conform to predefined formats. The biggest advantage of using EpiCaster is that it reduces the timeline for presenting results of forecasts on the Web for any new disease, or an existing disease like Flu in a new region. Using EpiCaster, the time for implementing such extensions can be reduced from several months of application development time to a handful of days of data manipulations.

Acknowledgments

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

References

1. Columbia Prediction of Infectious Diseases. <http://cpid.iri.columbia.edu/flu.html>
2. HealthMap. <https://healthmap.org/>
3. NDSSL: Ebola Research. <http://www.vbi.vt.edu/ndssl/ebola>
4. Barrett C, Beckman R, Khan M, Kumar V, Marathe M, Stretz P, Dutta T, Lewis B. Generation and analysis of large synthetic social contact networks. Proceedings of the Winter Simulation Conference (WSC). 2009:1003–1014.
5. Bisset K, Feng X, Marathe M, Yardi S. Modeling interaction between individuals, social networks and public policy to support public health epidemiology. Proceedings of the Winter Simulation Conference (WSC). 2009:2020–2031.

6. Bisset KR, Chen J, Feng X, Kumar VSA, Marathe MV. EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. *Proceedings of the 23rd International Conference on Supercomputing*. Jun.2009 :430–439.
7. Bisset, KR.; Chen, J.; Feng, X.; Ma, Y.; Marathe, MV. *Proceedings of the 24th ACM International Conference on Supercomputing, ICS '10*. New York, NY, USA: ACM; 2010. Indemics: An interactive data intensive framework for high performance epidemic simulation; p. 233-242.
8. Butler D. When Google got flu wrong. *Nature*. 2013; 494(7436):155. [PubMed: 23407515]
9. Chakraborty P, Khadivi P, Lewis B, Mahendiran A, Chen J, Butler P, Nsoesie EO, Mekaru SR, Brownstein JS, Marathe M, et al. Forecasting a moving target: Ensemble models for ili case count predictions. *SIAM International Conference on Data Mining*. 2014
10. Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, Toroczkai Z, Wang N. Modelling disease outbreaks in realistic urban social networks. *Nature*. 2004; 429:180–184. [PubMed: 15141212]
11. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2008; 457(7232):1012–1014.
12. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, Del Valle SY. Forecasting the 2013–2014 influenza season using wikipedia. 2014 arXiv preprint arXiv: 1410.7716.
13. Kallus N. Predicting crowd behavior with big public data. *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 625–630. International World Wide Web Conferences Steering Committee. 2014
14. Lofgren ET, Halloran ME, Rivers CM, Drake JM, Porco TC, Lewis B, Yang W, Vespignani A, Shaman J, Eisenberg JN, et al. Opinion: Mathematical models: A key tool for outbreak response. *Proceedings of the National Academy of Sciences*. 2014; 111(51):18095–18096.
15. Merler S, Ajelli M, Fumanelli L, Gomes MF, Piontti APy, Rossi L, Chao DL, Longini IM, Halloran ME, Vespignani A. Spatiotemporal spread of the 2014 outbreak of ebola virus disease in liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. *The Lancet Infectious Diseases*. 2015
16. Nsoesie E, Marathe M, Brownstein J. Forecasting peaks of seasonal influenza epidemics. *PLoS currents*. 2013; 5
17. Nsoesie EO, Beckman RJ, Shashaani S, Nagaraj KS, Marathe MV. A simulation optimization approach to epidemic forecasting. *PLoS ONE*. 2013; 8(6):e67164. 06. [PubMed: 23826222]
18. Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe MV. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and Other Respiratory Viruses*. 2013
19. Osgood, N.; Liu, J. *Proceedings of the 2014 Winter Simulation Conference, WSC '14*. Piscataway, NJ, USA: IEEE Press; 2014. Towards closed loop modeling: Evaluating the prospects for creating recurrently regrounded aggregate simulation models using particle filtering; p. 829-841.
20. Paul MJ, Dredze M. You are what you tweet: Analyzing twitter for public health. *ICWSM*. 2011:265–272.
21. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. *PLoS currents*. 2013; 6
22. Pigott DM, Golding N, Mylne A, Huang Z, Henry AJ, Weiss DJ, Brady OJ, Kraemer MU, Smith DL, Moyes CL, et al. Mapping the zoonotic niche of ebola virus disease in Africa. *Elife*. 2014; 3:e04395. [PubMed: 25201877]
23. Ramakrishnan, N.; Butler, P.; Muthiah, S.; Self, N.; Khandpur, R.; Saraf, P.; Wang, W.; Cadena, J.; Vullikanti, A.; Korkmaz, G., et al. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM; 2014. Beating the news with embers: forecasting civil unrest using open source indicators; p. 1799-1808.
24. Rivers C. Ebola: models do more than forecast. *Nature*. 2014; 515(7528):492–492.
25. Rivers C, Lofgren E, Marathe M, Eubank S, Lewis B. Modeling the Impact of Interventions on an Epidemic of Ebola in Sierra Leone and Liberia. *PLoS Currents*. 2014
26. Rivers CM, Lofgren ET, Marathe M, Eubank S, Lewis BL. Modeling the impact of interventions on an epidemic of ebola in sierra leone and liberia. 2014 arXiv preprint arXiv:1409.4607.

27. Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*. 2012; 109(50):20425–20430.
28. Shaman J, Yang W, Kandula S. Inference and forecast of the current west african ebola outbreak in Guinea, sierra leone and liberia. *PLoS currents*. 2014; 6
29. Sugumaran, R.; Voss, J. *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications*. ACM; 2012. Real-time spatio-temporal analysis of West Nile virus using twitter data; p. 39
30. Tamerius JD, Shaman J, Alonso WJ, Bloom-Feshbach K, Uejio CK, Comrie A, Viboud C. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS pathogens*. 2013; 9(3):e1003194. [PubMed: 23505366]
31. Tizzoni M, Bajardi P, Poletto C, Ramasco JJ, Balcan D, Gonçalves B, Perra N, Colizza V, Vespignani A. Real-time numerical forecast of global epidemic spreading: case study of 2009 a/h1n1pdm. *BMC medicine*. 2012; 10(1):165. [PubMed: 23237460]
32. Yang W, Karspeck A, Shaman J. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS computational biology*. 2014; 10(4):e1003583. [PubMed: 24762780]
33. Yang W, Lipsitch M, Shaman J. Inference of seasonal and pandemic influenza transmission dynamics. *Proceedings of the National Academy of Sciences*. 2015; 112(9):2723–2728.
34. Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS. Monitoring influenza epidemics in china with search query from baidu. *PLoS one*. 2013; 8(5):e64323. [PubMed: 23750192]

Figure 1.
Conceptual Architecture of EpiCaster

Figure 2.
Data Flow within EpiCaster

Figure 3.
Snapshot of EpiCaster as a standalone web application showing infection counts of Ebola in Bong county of Liberia

Figure 4. Forecasting of Ebola in Sierra Leone without considering a change in the transmissibility observed through surveillance

Figure 5. Forecasting of Ebola in Sierra Leone considering a only one transition in transmissibility observed through surveillance

Figure 6. Forecasting of Ebola in Sierra Leone considering both transitions in the transmissibility observed through surveillance

Table 1

Response times and throughput of the API – *getOverallInfection()* when 100 users access the API over the period of 1 to 20 seconds

Ramp-up time for 100 users	Avg Response Time	Median Response Time	Throughput (requests/sec)
20 sec	53 ms	28 ms	19.9
10 sec	99 ms	33 ms	10.1
5 sec	199 ms	35 ms	19.7
2 sec	1515 ms	1511 ms	38.9
1 sec	1952 ms	1928 ms	39.8

Table 2

Response times and throughput of 2 different APIs accessed with different regions and diseases

API Name	Region	Avg Response Time	Median Response Time	Throughput(requests/second)
<i>getInfectionPlot()</i>	Virginia	88 ms	34 ms	10
	Wyoming	90 ms	45 ms	10
	California	99 ms	33 ms	10.1
	Delhi	57 ms	33 ms	1.2
	Liberia	89 ms	31 ms	10.1
<i>getTop3InfRegions()</i>	Virginia	154 ms	132 ms	11.2
	Wyoming	254 ms	156 ms	10
	California	190 ms	133 ms	10.1
	Delhi	111 ms	54 ms	10
	Liberia	76 ms	44 ms	10