Procedia Computer Science

Volume 51, 2015, Pages 765–774

ICCS 2015 International Conference On Computational Science

Mining Mobile Datasets to Enable the Fine-Grained Stochastic Simulation of Ebola Diffusion

Nicholas Vogel[1], Christopher Theisen[1], Jonathan P. Leidig[1], Jerry Scripps[1], Douglas H. Graham[2], and Greg Wolffe[1]

[1] School of Computing and Information Systems, GVSU, Allendale, MI
[2] Department of Biomedical Sciences, GVSU, Allendale, MI

Abstract

The emergence of Ebola in West Africa is of worldwide public health concern. Successful mitigation of epidemics requires coordinated, well-planned intervention strategies that are specific to the pathogen, transmission modality, population, and available resources. Modeling and simulation in the field of computational epidemiology provides predictions of expected outcomes that are used by public policy planners in setting response strategies.

Developing up to date models of population structures, daily activities, and movement has proven challenging for developing countries due to limited governmental resources. Recent collaborations (in 2012 and 2014) with telecom providers have given public health researchers access to Big Data needed to build high-fidelity models. Researchers now have access to billions of anonymized, detailed call data records (CDR) of mobile devices for several West African countries. In addition to official census records, these CDR datasets provide insights into the actual population locations, densities, movement, travel patterns, and migration in hard to reach areas. These datasets allow for the construction of population, activity, and movement models. For the first time, these models provide computational support of health related decision making in these developing areas (via simulation-based studies).

New models, datasets, and simulation software were produced to assist in mitigating the continuing outbreak of Ebola. Existing models of disease characteristics, propagation, and progression were updated for the current circulating strain of Ebola. The simulation process required the interactions of multi-scale models, including viral loads (at the cellular level), disease progression (at the individual person level), disease propagation (at the workplace and family level), societal changes in migration and travel movements (at the population level), and mitigating interventions (at the abstract governmental policy level). The predictive results from this system were validated against results from the CDC's high-level predictions.

Keywords: Agent-based Modeling, Ebola, Stochastic Simulation

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.197

1 Introduction

The cost of infectious diseases has a significant negative impact on the economy, health, and well-being of countries. The cost of succumbing to and recovering from infectious diseases often places the highest burden on poor and disadvantaged citizens. These individuals are unable to acquire sufficient preventative, diagnostic, and treatment services. In addition, the effects of an infection are generally more severe in high-risk groups, e.g., pregnant, HIV-positive, and the elderly. Thus, it is especially important to maximize the use of public resources and optimize policies related to health. In West Africa, communicable diseases (e.g., Ebola and meningitis), vector-borne diseases (e.g., malaria and yellow fever), and parasitic diseases (e.g., Schistosomiasis) have a significant impact on health and the economy. Senegal in particular is at high-risk for infectious diseases and epidemics. Medical resources (such as physicians and pharmaceutical treatments) and infrastructure (hospitals and clinics) are limited in many developing areas despite the prevalence of infectious diseases. Thus, it is not straightforward to efficiently prevent and respond to epidemics or eradicate diseases. Proper planning is required to best allocate and use available resources, especially if an epidemic threatens to exhaust resource availability. Governmental policies are required for activities such as closing borders, closing schools, stopping commerce, conducting surveillance, outreaching through mass media, distributing pharmaceutical treatments, restricting travel, performing quarantine, and isolating. However, it is not known a priori which optimal combination of mitigation strategies will best prevent or end an epidemic. Simulation results provide a computational prediction for how a given disease will spread in a given population and scenario. Conducting a simulation study allows governmental officials to set policy based on predicted future events, infections, and cost.

Several simulation software tools have been developed to simulate the spread of diseases [2, 3, 7, 12]. These computational epidemiology tools have been used extensively in setting public health policies for several national governments and aid organizations. The tools were designed to handle a range of diseases including avian flu, pertussis, smallpox, and malaria. However, the current generation of software tools must be modified to accurately simulate the spread of Ebola. These tools have been used to set policies in several developed countries with population models built on government census records and individual activity questionnaires. The tools require models for populations, social networks, individual behavior, movement, and diseases. However, these models have not been developed or parameterized to work well with developing countries and the Ebola virus. Many assumptions and population models produced for developed countries break down when applied to the spread of diseases in West Africa. There are differences in family and household sizes, age structures, school sizes and attendees, lifestyles, social networks, movement models, migration, seasonal shifts, climate, transportation infrastructure, and the locations and availability of healthcare resources. Limited government resources and difficulty of travel in certain areas have prevented some countries from producing censuses and activity models. With the scarcity of information regarding remote areas, it has not historically been possible to construct accurate models regarding health in these regions.

Fortunately, mobile phones are ubiquitous in developing countries. Anonymous call detail records (CDR) provide metadata regarding the time and location a person sends or receives a call or SMS text message. With anonymized CDR datasets, researchers are able to track relative population levels in each area of the country, individual movements, seasonal hotspots, population shifts, and migration. With these datasets, data mining as applied to the frequency and timing of calls and texts make it possible to identify population trends. With recently made available CDR datasets, population and movement models can now be produced that enable more accurate simulations of epidemics in these areas. Population models of Ivory Coast and

Senegal now contain fine-grained detail on population travel, daily movement, and interactions.

Governmental public health officials may utilize the newly developed models and modified simulation software to set public policy in West Africa. This stochastic approach has been benchmarked and calibrated against predictions from analytical CDC models.

2 Computational Epidemiology and Policy

Stochastic models are widely used in public health research, predicting a variety of scenarios from propagation (e.g., HIV spread in prisons [13]) to treatment (e.g., optimizing Emergency Department organization [4]). Modeling and simulation software provides healthcare planners with the ability to predict the results of scenarios. These scenarios consist of a hypothetical situation consisting of hundreds of variables. The population for the simulation may consist of a village, arrondissement, department, nation, or set of countries. Research groups have produced software applications for computational experimentation related to the spread of diseases. These tools typically require population models, activity models, and disease models. Population models include individuals' demographics, population age structures, and population densities. Activity models describe travel patterns, daily schedules, and interactions between individuals. Disease models required for these simulation applications have been developed during research concerning the spread of diseases in other countries. However, disease models for Ebola have not gained widespread adoption and use in computational epidemiology due to the slow emergence of concrete facts regarding the current outbreak. With these models and datasets, a software application may predict the diffusion of Ebola throughout the population. Stochastic software models are used to predict the probabilistic spread between individual hosts.

Public health mitigation strategies play a major role in preventing an epidemic and facilitating its eradication. Experimental studies in the field of computational epidemiology are conducted by executing the simulation software using the underlying models and datasets in order to compare the results of thousands of scenarios. These results then lead to best practices.

3 Model Creation for Stochastic Simulations

The field of computational epidemiology makes use of computing to improve population health by informing public policy. Distributed, stochastic simulation software was developed to forecast the spread of Ebola in West Africa given a potential scenario. The software requires models of the Ebola virus, West Africa populations, mobility, mixing patterns, and government policy.

3.1 Mobile Datasets

Access to CDR data provided by Orange Telecom and the Senegal and Ivory Coast governments through the D4D Challenge allowed the development of population models previously unavailable to stochastic, agent-based simulation platforms. Active user ids, defined as having at least one CDR record, were used to provide details of human mobility and locations based on arrondissements and antennas. The antenna-based datasets consist of 25 two-week duration records each with 300,000 individuals and overlap of some individuals between records. Population modeling and knowledge discovery based on these datasets makes use of the latent information contained therein. The continuous monitoring and geospatial details provide higher resolution data than previously available in census records. Constructing population models enables research in health and disease propagation and provides the basis for stochastic models.

In order to utilize the 1,204,451,385 datapoints in the fine-grained dataset, multiple programs were developed to mine CDRs and identified useful information. User ids and timestamps were recorded to determine the number of calls per user at a sampling site. A site may be a geographical region (prefecture or arrondissement) or antenna. The average total calls/texts by a user for the entire year was 3,764, or roughly 10 calls/texts per user per day. From these records, a user's home and work locations were matched with antenna locations based on time of day. To ascertain human mobility data, the CDRs were mined to track movements between antennas in chronological order by a user id. This was accomplished by ordering the sequential records of a user's location by timestamp and then determining the time periods when they were observed moving to a different antenna location. Subsequent sites were only recorded if the antenna location was different from the previously recorded location. In order to determine the contribution of a particular population to overall population movement in a macro view, 'hops' were defined as a user traveling to a new antenna location. Hop width is the accumulation of the movement for the whole population between each pair of sites. In the following mobility figures, each time a user travels to a different site, the thickness of the line between the previous and current sites increases, indicating increased user movement between the sites. This process produced D4D-informed population and mobility models, discussed in the following sections.

3.2 D4D-Informed Synthetic Populations

When synthesizing Senegal's population, the ratio of the percentage of mobile users per geographic antenna location, derived from CDR dataset, was used to scale the spatial population density for over 1,600 antenna locations. By multiplying Senegalese census regional population data by this ratio, the population was distributed with fine-grained detail to each antenna. The population's age distribution was applied to each area and antenna range. Using the cumulative ratios of age distribution, ages were assigned to each individual. Household size distributions were constructed as an array using Senegalese household size data [10]. Using the cumulative ratios of household sizes, the size of each household was probabilistically assigned. In constructing the set of individuals occupying a household, an adult of working age was first added to the household (either aged 15-24 or 25-64), with associated probabilities based on population age distribution. If the household size was larger than one, the other individuals were randomly assigned according to age distribution probabilities of the population. Figure 1 provides an overview of population modeling given the antenna-based sampling locations. This process led to the construction of a synthetic population model that contains an individual agent for each resident, placed in locations and family units statistically aligned with observed mobile records.

3.3 D4D-Informed Mobility and Activity Modeling

The population, worker-flow, and employment models specify where every individual lives, works, and travels based on their observed movement. Each individual's home and work locations were calculated based on the most frequently used antenna between the times of 7:00pm-7:00am (home) and 7:00am-7:00pm (work). Home locations are displayed in Figure 1 and align with census estimates in each geographic region. A similar work location for each individual was assigned from CDR and employment data. The employment data, such as working age population and percentage of employment, is used to model the percent of employed working age individuals who travel to non-home locations for work. The antenna-based work population (7:00am-7:00pm) was scaled by the percent of working age individuals and the percent of employment in each area in order to probabilistically assign employment to individuals at

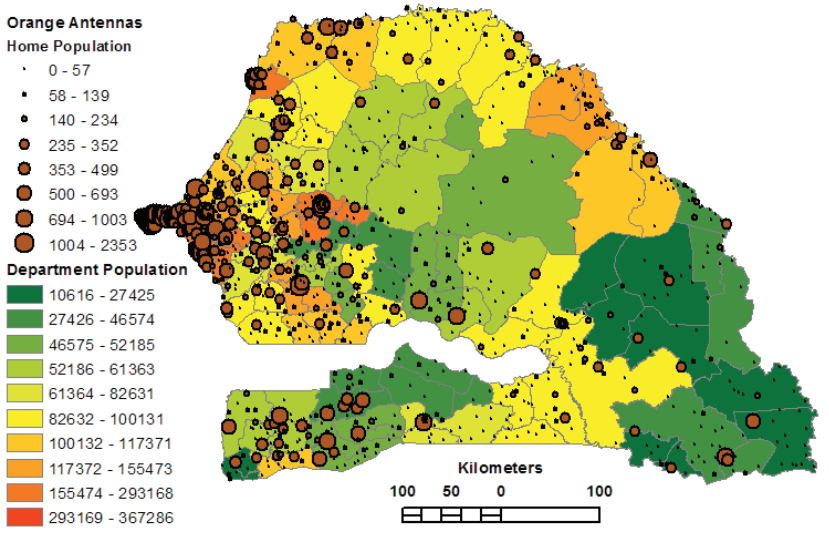

Figure 1: Population density, sampling locations (antennas), and distribution of home locations in Senegal based on sampling at antenna locations in the D4D datasets. The size of each political department is also provided from a recent census.

each antenna location. This brings workers into contact (through the daytime movement of individuals) who are located at work within the same antenna range. Figure 2 displays several geospatial factors in the movement of people and spread of Ebola in Senegal, such as roadways, border crossings, and location sampling (i.e., antenna locations). For the capital of Dakar, this figure highlights the locations with increased social mixing through daytime interactions and work locations. This mobility model improves the accuracy of simulating social connections and mixing patterns between agents. This was accomplished by using the datasets of mobility traces for hundreds of thousands of random individuals located throughout the country. Figure 3 displays the total travel between political districts within Senegal, i.e., arrondissements. Previously, simulations of remote or developing areas relied on coarse-grain, fully mixing models that were produced from the same assumptions made for developed countries with completely different social patterns or else extrapolated from limited, small scale surveys conducted by workers on the ground. Through the D4D datasets, it is now possible to assign mobility based on actual observed behavior, notably in remote locations.

3.4 Viral Trajectory, Disease Progression, and Propagation

The Ebola virus causes an often-fatal disease, with major outbreaks occurring over the last 40 years in Sudan, the Democratic Republic of Congo, and now across West Africa (Sierra Leone, Liberia, Guinea, Nigeria, Mali, and Senegal). It is spread through contact with infected animals (e.g., bats) or human bodily fluids. Modeling an Ebola epidemic requires models of propagation between hosts and progression within a host. Computational epidemiology models typically maintain a finite-state model for each individual. Each agent in the population maintains a current state: susceptible, incubating, infectious, and recovered/fatality (i.e., an SIIR model). Infected hosts follow one of two trends (viral load trajectories) that guide viral replication and disease progression. The predicted daily viral load is used to provide realistic estimates

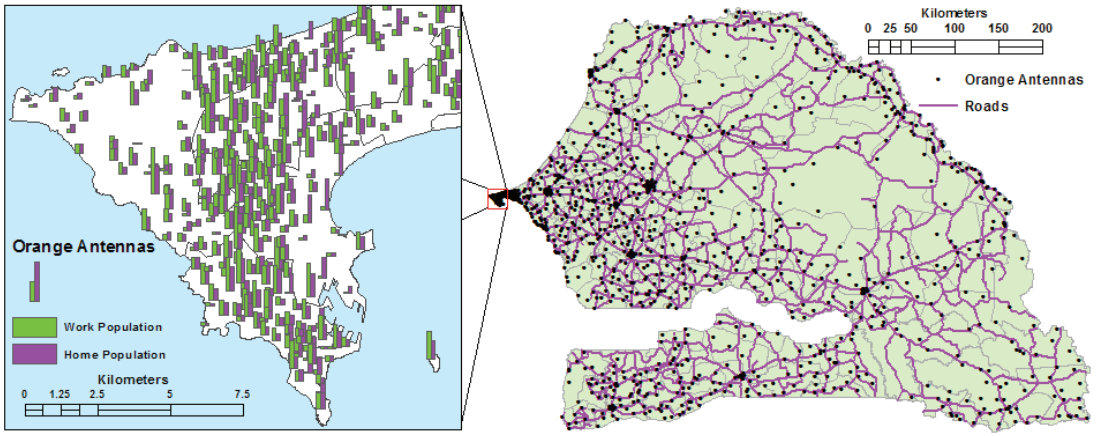

Figure 2: Daily net migration (left callout) between residential areas and economic centers (e.g., industrial, commercial, and agricultural) based on antenna locations in the capital of Dakar. Geospatial map (right) of the roads, border crossings, and sampling locations (antennas). Movement between the sampling locations provided high-resolution detail of mobility, social contacts, and disease transmission pathways in the population models.

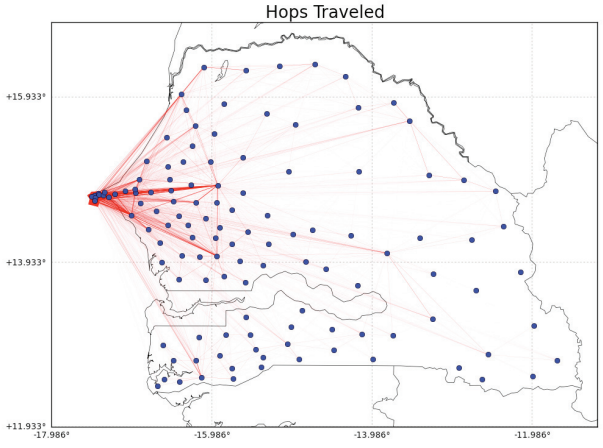

Figure 3: Mobility traces (travel) between arrondissements. The thickness of an edge between two arrondissements represents the number of times individuals were observed traveling directly from one arrondissement to the other, in either direction. A round trip between the two locations would count as two separate hops.

of the first onset of symptoms and first possible diagnosis after the host is infected. The time range between infection and first possible diagnosis (ascertainment delay) represents the earliest length of time in which Ebola is detectable with current lab procedures [6]. Two viral load trajectories (which determine the extent to which the virus replicates within a host) were

developed, one resulting in death. The trajectories were based upon historical fatality rates ranging from 40-70% [5] and prior determined trajectories [14].

3.5 Simulation Details

The software produced in this research was modified from an open-source software package, FluTE. FluTE is a stochastic, agent-based simulation system for modeling the spread of influenza, based upon United States census data [7]. The program was modified and rewritten to model the direct-contact transmission of the Ebola virus within the Senegalese population. The process for producing a simulation engine for Ebola in Senegal required three steps.

- A population model was produced using D4D Challenge and census data. The developed model assigns the home and travel locations of all 13 million residents to antenna locations.
- A mobility model for travel and movement within Senegal was produced using the D4D dataset. The developed model contains the day and nighttime travel patterns of all residents based on the observed movement between antennas within the dataset.
- FluTE's source code was modified to provide a platform for Ebola progression and propagation as well as Ebola transmission related parameters.

The simulation system requires four datasets as input to produce a prediction: geo-political models (location of antennas and arrondissements), worker movement data (travel between locations), employment data (movement of workers), and scenario configuration files. These input datasets to the simulation system are based on the population and mobility models, as described in sections 3.1-3.3. The following section summarizes the modifications of the existing computational epidemiology platform and discusses the limitations of the simulations.

3.6 Disease Modeling

Along with the construction of D4D-informed population models, the simulation software was modified to support prediction of Ebola. Table 1 details several modifications required to properly describe Ebola. In contrast with the standard influenza duration period of a few days, the incubation period for Ebola is many times longer. The incubation time period was changed to 2-21 days [5]. This delays the onset of the disease and leads to a slow growing epidemic. While influenza outbreaks are expected to be seasonal, Ebola outbreaks require more computational processing due to the longer simulation time required for a multi-year simulation.

Parameter	Range
Incubation period	2-21 days
Simulation length	Multiple seasons and years
Viral Load Trajectories	2 trajectories, one fatal
Case fatality rate	40-70%
Days after death before burial	1-3 days

Table 1: Software modifications required to simulate Ebola.

Some features required to simulate Ebola already existed as parameters in FluTE. However, minor software updates and scenario file parameterizations were required. The two possible viral load trajectories of a patient were linked to the fatality of their Ebola case. As FluTE does not simulate death, fatality was added in order to investigate the effects of burial practices.

Parameter	Range
Reproduction rate (R_0)	1.51-2.53
Ascertainment delay	3 days after symptomatic
Models	Population, airports, ports, borders, political boundaries, mobility

Table 2: Configuration parameters required to simulate Ebola.

Burial rituals were added to the simulation platform based on cultural practices and may be modified through intervention policies. After the person is deceased, they are isolated at home for 1 to 3 days to simulate burial practices in which family members prepare the deceased person before burying them [9]. During this time, the deceased individual has their viral load set to the highest level within their trajectory to simulate the high viral loads seen in patients before and after dying from Ebola [8]. After the funeral, the deceased is removed from the population.

The basic reproduction rate in Table 2, denoted R_0 (the number of secondary infections generated from an infectious person) was changed to the range 1.51-2.53 [1]. The ascertainment delay is based on the time required to diagnosis Ebola in a patient using current experimental lab techniques [6]. Multiple parameters are used to implement governmental policies, e.g., closing schools and borders. The software allows seeding infected persons in major ingresses, such as airports. Thus, geospatial datasets regarding the location of airports, ports, and boarder crossings were added to the software's underlying assumptions. In addition, the software includes features such as mitigation measures to reduce the extent of an outbreak.

There were aspects of FluTE that were not applicable to the Ebola disease model, such as antiviral kits and vaccinations. Although research continues in this area, neither pharmaceutical intervention is available for mitigation efforts as of this publication. Temporal seasonality was inactivated, as historical outbreaks of Ebola do not correlate to specific times of the year [11].

4 Model Validation

The simulation software provides support for public health policy regarding the spread of Ebola given potential scenarios. The developed model required extensive modifications of the FluTE platform. To verify the results of the system, worst-case scenarios were compared with output from a publicly available model provided by the Centers for Disease Control and Prevention [15]. The CDC's predictive model provides a coarse-grained extrapolation of expected infection counts. The CDC model predicts daily Ebola infections for a generic population of a user defined size. However, the analytical model does not take the population structure, dynamics, or topology into consideration or allow for predicting the effects of mitigation strategies. The model provides a worst-case expectation for the spread of Ebola in the absence of public policy or basic actions taken by individuals (e.g., staying home when sick). Figure 4 displays the results of the D4D-informed stochastic model. For this figure, simulations were run with the Senegalese population of 13,401,076, one initial index case, and parameters encoding current assumptions regarding the characteristics of Ebola. The D4D-informed model produces results in alignment with CDC predictions in addition to providing finer-resolution information regarding each infected individual. As shown in Table 3, the CDC model predicts between 6.9 and 11.9 million infections depending on the average number of days a person is infectious. This is in alignment with the stochastic model's prediction of 8.9 million infections. The close alignment demonstrates that the disaggregated, stochastic Ebola simulation model produces the type of predicted result expected by analytical governmental agencies using high-level dif-

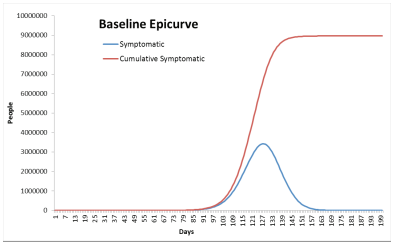

Figure 4: Simulation results from the D4D-informed model detailing the count of currently infected individuals (symptomatic) and total that have been infected (cumulative symptomatic). These sample results are for the worst-case baseline scenario without any governmental intervention strategies.

Number of days infectious	Total cases
10 (CDC Model)	6,886,748
11 (CDC Model)	9,370,528
12 (CDC Model)	10,921,851
13 (CDC Model)	11,939,709
Average result (our model)	8,966,151

Table 3: Predicted attack rates for a year-long simulation by the analytical CDC model and stochastic D4D-based model in a baseline scenario (i.e., no governmental interventions or preventative actions taken by individuals). The new stochastic model may be used to predict the infection reduction after applying optional governmental intervention strategies.

ferential equations. The D4D-based population model provided a realistic dataset that led to the aligned simulation results. Senegal's single confirmed Ebola case (fatal) in 2014 involved a sick individual arriving in Dakar from neighboring Guinea. He was isolated by the Senegalese government and did not spread the disease further even after 67 contacts with family and health workers. Mobile datasets from neighboring countries would lead to additional calibration and validation. In addition, future studies may predict the epidemics that result from policies.

5 Summary

Public health policies regarding preparation and planning for Ebola epidemics require accurate predictions for successful mitigation and optimal use of resources. Computational epidemiology and bioinformatics modeling of Ebola were developed to inform the simulation system regarding viral progression within a host and the transmission between hosts. The simulation software was developed to predict the spread of Ebola, specifically within West African countries. With the software, public health officials may evaluate the cost, effect, and success of potential mitigation strategies. The interventions available to policy makers include quarantine, isolation, border closings, public and economic closures, travel restrictions, etc. Pharmaceutical interventions (vaccines and antiviral kits) are possible to simulate should these treatments become available.

Stochastic models provide the possibility to analyze potential mitigation strategies in order to set optimal public health policies. The use of our population, mobility, and stochastic simulation models provide more accurate simulation details in comparison to high-level analytical predictions. The D4D mobile datasets provide high-resolution information useful for modeling developing regions and hard to reach locations. The population models are based on the latent movement, mobility, and population densities in Senegal and Ivory Coast. The simulation software was shown to provide results in alignment with CDC Ebola predictions.

Acknowledgments

We thank France Telecom-Orange and the D4D Challenge for providing access to Cote d'Ivoire and Senegal mobile datasets. Special thanks to Kurt O'Hearn for comments and suggestions.

References

[1] Christian L. Althaus. Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. In *PLoS Currents: Outbreaks*, September 2014.

[2] Christopher L. Barrett, Keith R. Bisset, Stephen G. Eubank, Xizhou Feng, and Madhav Marathe. EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *SC '08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, pages 1–12, Piscataway, NJ, USA, 2008. IEEE Press.

[3] Keith Bisset, Jiangzhuo Chen, Xizhou Feng, Anil Vullikanti, and Madhav Marathe. EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *ICS '09: Proceedings of the 23rd International Conference on Supercomputing*, pages 430–439, New York, NY, USA, 2009. ACM.

[4] Eduardo Cabrera, Manel Taboada, Ma Luisa Iglesias, Francisco Epelde, and Emilio Luque. Simulation Optimization for Healthcare Emergency Departments. *Proceedings of the International Conference on Computational Science (ICCS)*, 9(0):1464 – 1473, 2012.

[5] CDC. Ebola outbreak in West Africa case counts. `http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/case-counts.html`, December 2014.

[6] CDC. Interim guidance for specimen collection, transport, testing, and submission for patients with suspected infection with Ebola virus disease. `http://www.cdc.gov/vhf/ebola/pdf/ebola-lab-guidance.pdf`, December 2014.

[7] Dennis L. Chao, M. Elizabeth Halloran, Valeria J. Obenchain, and Ira M. Longini. FluTE, a Publicly Available Stochastic Influenza Epidemic Simulation Model. *PLoS Comput Biol*, 6(1), 2010.

[8] Scott F. Dowell, Rose Mukunu, Thomas G. Ksiazek, Ali S. Khan, Pierre E. Rollin, and Clarence J. Peters. Transmission of Ebola hemorrhagic fever: a study of risk factors in family members, Kikwit, Democratic Republic of the Congo, 1995. *Journal of Infectious Diseases*, 179:87–91, 1995.

[9] Jean-Claude Legrand, Rebecca F. Grais, Pierre-Yves Boelle, and Alain-Jacques Valleron. Understanding the dynamics of Ebola epidemics. *Epidemiology and Infection*, 135(4):610–621, 2007.

[10] Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 6.3 [Machine-readable database]. *Minneapolis: University of Minnesota Technical Report*, 2014.

[11] Sophia Ng and Ben J. Cowling. Association between temperature, humidity and ebolavirus disease outbreaks in Africa, 1976 to 2014. *Euro Surveillance*, 19(35), 2014.

[12] Thomas Smith, Gerry F. Killeen, Nicolas Maire, Amanda Ross, Louis Molineaux, Fabrizio Tediosi, Guy Hutton, Jurg Utzinger, Klaus Dietz, and Marcel Tanner. Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of Plasmodium falciparum malaria: Overview. *Am J Trop Med Hyg*, 75:1–10, 2006.

[13] Alfredo Tirado-Ramos and Chris Kelley. Next Steps in Simulating High-risk Infectious Disease Propagation Networks. *Proceedings of the International Conference on Computational Science (ICCS)*, 18(0):1421 – 1428, 2013.

[14] Jonathan S. Towner, Pierre E. Rollin, Daniel G. Bausch, Anthony Sanchez, Sharon M. Crary, Martin Vincent, William F. Lee, Christina F. Spiropoulou, Thomas G. Ksiazek, Mathew Lukwiya, Felix Kaducu, Robert Downing, and Stuart T. Nichol. Rapid diagnosis of Ebola hemorrhagic fever by reverse transcription PCR in an outbreak setting and assessment of patient viral load as a predictor of outcome. *J Virol*, 78:4330–4341, 2004.

[15] Michael Washington, Charisma Atkins, and Martin Meltzer. CDC Generic EbolaResponse (ER), Modeling the spread of disease impact and intervention, version 2.5. `http://stacks.cdc.gov/view/cdc/24900`, September 2014.