Predicting the 2014 Ebola Outbreak in West Africa using Network Analysis

Milestone Report

Shafi Bashar, Mike Percy, Romit Singhai

{*shafiab, mp81, romit*}*@stanford.edu*

Abstract

The current Ebola outbreak in West Africa is the worst in history. Most traditional epidemiological models are compartmental models that have a random-mixing assumption, calculating the effective reproductive rate of an outbreak. We survey three of these models: the classic SIR (Susceptible, Infectious, Recovered) model and two extensions used in Ebola research.

Network models allow for avoiding the random-mixing assumption inherent in compartmental models. This is done by assigning each individual a finite set of permanent contacts. We review generated contact network models for SARS, including an urban network, a random network, and a scale-free network. We then propose a worldwide network model that takes an alternate approach to modeling the flow of people based on economic trade data and attempt to predict the future of the Ebola outbreak.

I. Introduction

In this paper we explore predicting an Ebola outbreak using a worldwide network of trading countries.

II. Related Work

Related work by (**?**) attempts to predict the spread of Ebola to different parts of the world based on a model that incorporates both the compartmental approach and the use of world-wide air traffic flows.

In (**?**), the authors model the spread of the 2002-2003 outbreak of SARS in Hong Kong and Canada using a contact network. A contact network model attempts to characterize every interpersonal contact that can potentially lead to disease transmission in the community, with each person in the community represented as a node and each contact represented as an edge between them.

The majority of research in epidemiological theory is based on the compartmental model, which is not a network model. In order to simply capture the dynamics of disease spread over time, the compartmental model employs a a population-wide random mixing assumption, meaning that each individual has a small and equal chance of coming into contact with any other individual in the population. To model the progress of an epidemic in a large population, the individuals in the population are compartmentalized according to the state of the disease. The most widely used such model is the SIR model introduced in (**?**):

- **Susceptible (S):** Individuals who have not yet caught the disease from contact with an infectious individual.
- **Infectious (I):** Individuals who have the disease. They have some probability of infecting susceptible people.
- **Recovered (R):** Individuals who have experienced the full infectious period, and are now non-infectious and immune.

The changes among these states over time are represented by a set of differential equations. The basic reproductive number R_0 is defined as the average number of secondary cases generated by a primary case in a pool of mostly susceptible individuals, and is an estimate of epidemic growth at the start of an outbreak if everyone is susceptible.

In (**?**), the authors model the effect of Ebola outbreaks in 1995 in Congo and in 2000 in Uganda using a compartmental model similar to the SIR model. However, a distinct feature of Ebola is that individuals exposed to the virus who become infectious do so after a mean incubation period. In order to reflect this feature, in the SIR model is extended with an additional "Exposed" compartment state. This SEIR model is summarized in Figure **??**:

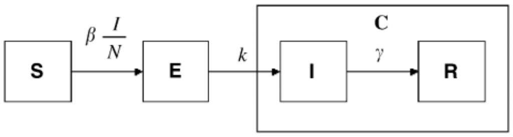


Fig. 1: SEIR model

In (**?**), the Ebola outbreaks in Congo in 1995 and Uganda in 2000 are also studied. However, they expand on the model from (**?**) by modeling the spreading of Ebola in heterogeneous settings. In order to gain more insight into the epidemic dynamics, the infectious phase is subdivided into three stages: infection in a community setting (I), infection in a hospital setting (H), and infection after death assuming a traditional funeral (F). The resulting model is known as the SEIHFR model. While this model provides finer grained modeling of the behavior of Ebola, it is also complex and the link to a network-based model becomes tenuous. Because of this, we do not plan to use the SEIHFR model in our work.

III. Modeling localized epidemic spread

Estimating model parameters and basic reproduction number for 2014 Ebola data using a random mixing model

In first phase of the project, to calculate the basic reproductive number of the current Ebola epidemic spread in different countries, we performed model fitting on the data we gathered from **?**. We considered the SEIR model described in **?**. The SEIR model under consideration is a non-linear model with six parameters. The current Ebola epidemic is still spreading and depending on the preventative measures taken, the underlying dynamics of the spread can change drastically any time.

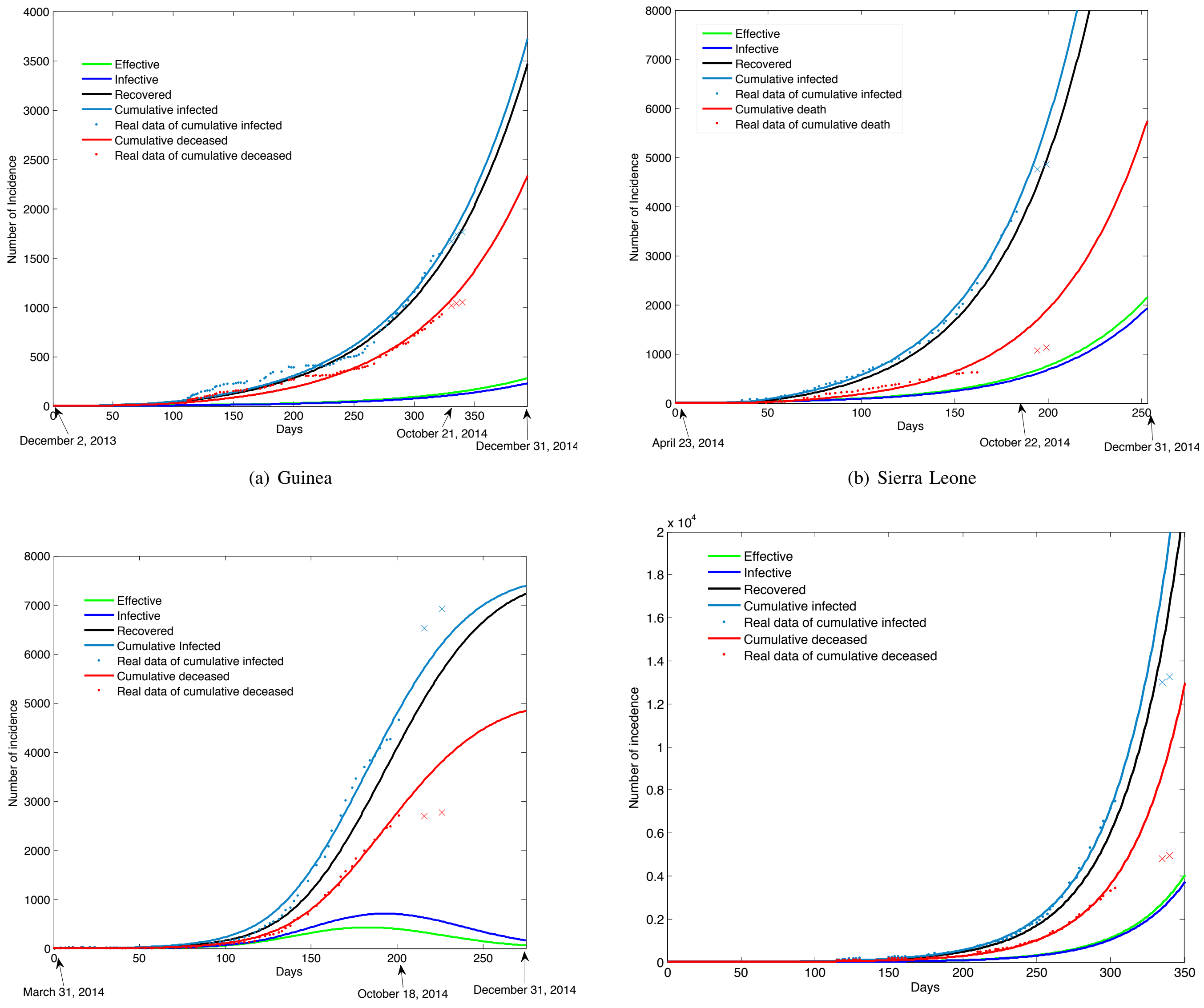
We fitted our model to three countries in West Africa: Guinea, Sierra Leone and Liberia. In addition, we performed model fitting for the West Africa region by adding up the data from these three countries. Given the limited number of data available, instead of fitting all six parameters to the model, we decided to fix some of the parameters based on studies on previous Ebola epidemics. In **?**, the incubation time of the Ebola virus $1/k$ is found to vary between 1 and 21 days, with a mean time of 6.3 days for previous Ebola outbreaks. For ease of data fitting, we set this parameter value to the mean value of 6.3 days. We note that the dynamics of the current epidemic may differ from previous ones, and therefore fixing a value based on the prior estimate may lead to some inaccuracy. To model the effect of intervention on the spread of the epidemic, a modified transmission rate β_1 is generally used instead of the initial transmission rate β_0 at later stages of the epidemic. The transition of β_0 to β_1 depends on intervention time τ and decaying factor q. The choice of intervention time is a very difficult problem. To this end, we looked into different sources like Wikipedia, WHO and CDC website to learn more about the timeline of the spread. In Guinea, a 2-year-old boy died on December 2, 2013, later diagnosed as an Ebola patient. We consider this incident as the index case for Guinea and set t_0 to December 2. On March 2, the Government of Guinea informed WHO regarding the possibility of an Ebola epidemic and declared a national heath emergency. We considered this date as the intervention date and set τ to 110. In Sierra Leone, one person died on April 2014. In June 12, 2014 the country declared an emergency and closed its borders with neighboring Guinea and Liberia. We consider the first date as t_0 and second date as the intervention time, therefore set τ to 50. In Liberia, on March 31, 2014, there was official confirmation of two people infected with Ebola. We set this date to t_0 and set I_0 and C_0 in our model as 2. The government of Liberia shut down all schools on July 30, 2014. We consider this date as the date of intervention in Liberia and set τ to 120.

In order to fit the non-linear SEIR model to the data, we used non-linear least squares estimation. We use the reported data (t_i, c_i) for $i = 1, 2, \ldots, n$ where t_i denotes the $i - th$ reporting time and c_i the cumulative number of infectious cases from the beginning of the outbreak at time t_0 to time t_i. The optimization problem contains a large number of local minima. Therefore, the choice of initial parameter estimate is an important consideration to get the globally optimal solution. In order to find a good initial choice of the parameters as input to the non-linear least squares solver, we first perform a Latin hypercube sampling on the 4-dimensional parameter space. We grid up the hypercube with a number of grid points in each dimension. We then choose the sample that minimizes the least squares error as the initial input. In order to calculate the 95% confidence interval of the estimated parameter, we performed bootstrapping based on residual error.

The estimated parameters for the SEIR model of Guinea, Sierra Leone, Liberia, and the overall West Africa region is presented in the Appendix. In Figure **??**, we presented the number of incidents at different compartments of the SEIR model, as well as the cumulative number of infectious cases over time. Our model fit was based on the last data we collected on October 21, 2014. As of today, November 11, 2014, few additional data points are available. We also plotted those data points on the graph. In addition to that, we extrapolated the graph up to December 31, 2014. We note that forecasting future cases may not be accurate as the underlying factors of the epidemic are changing rapidly with the increase in safety measures.

We observe that the prediction of the model to the cases up to November 11 is mostly on par with the observed data for Guinea and Sierra Leone. For the Guinea epidemic, we have data for the longest range of days as the initial case was on December 2. The estimated model, therefore, captures most of the dynamics of the spread in Guinea in contrast to the other cases. In case of Liberia, our prediction underestimated the observed data. Our guess is that the estimated parameters for Liberia may be overfitting the model. While performing some exploratory analysis on the Liberia data, we observed some discrepancies, e.g. decrease in cumulative value from the previous data points, which may somewhat distort the model fitting parameters. For the West Africa region plot, our prediction number is much higher than the actual data. This is understandable, as the underlying assumption behind the SEIR model is random mixing. Within a country without any movement restriction, this model is somewhat appropriate. However, for the aggregated model among multiple countries, this assumption is no longer valid due to stricter movement regulations between country borders. Therefore, a model with a random mixing assumption will overestimate the number of infectious cases.

As an extension of the above approach, for finding parameters for the model based on the current data, we designed and ran a simulation using an MCMC approach with input as time series data for the various infected regions. Initial values for the



(a) Guinea

(b) Sierra Leone

(c) Liberia

(d) West Africa

Fig. 2: SEIR model fit results for 2014 Ebola epidemic data

various parameters were assumed as above for the prior distribution, but a range based on the literature was used to find the most appropriate values for the current outbreak. For example β_0 had a range of $[0, 1]$, β_1 had a range of $[0, 1]$, infection time $1/\gamma$ had a range of $[3.5, 10.7]$, incubation time $1/k$ had a range of $[5, 22]$, and τ had a range of $[100, 150]$. This simulation is running and we will have results in the final report.

Mapping the compartmental model to the network based model

It is preferable to use a model that does not assume random mixing. One such network model we are integrating is a data set from a social networking site from the **?**. This undirected network has 4,846,609 nodes and 42,851,237 edges and average degree of 17.7. It is clearly a scale-free network because the degree distribution follows a power law after a degree of approximately 50. Using this data set, we calculate transmissibility and epidemic threshold.

We calculated the transmissibility T of a disease which is defined as the average probability that an infectious individual will transmit the disease to a susceptible individual with whom they have contact. We also calculated the epidemic threshold T_c which is the minimum transmissibility required for an outbreak to become a large-scale epidemic.

We start with the reproductive number R_0 calculated using the SEIR compartmental model. Using the relationship between the basic reproductive number R_0 and the transmissibility T given in **?**:

$$R_0 = T\frac{\langle k^2\rangle}{\langle k\rangle - 1}$$

while using values $R_0 = 2.27$ and $k = 17.7$ we find that transmissibility $T = 0.1210$ for this network.

We also calculate the value of epidemic threshold T_c, which is defined by Meyers as:

$$T_c = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}$$

for mean degree $\langle k \rangle$ and mean square degree $\langle k^2 \rangle$. Our calculations show that $T_c = 0.0598$ for this network.

We are using these results to predict average size of the outbreak and to compare the behavior of a random mixing compartmental model and a network model. We are incorporating this as a local model into our worldwide simulations.

IV. Predicting worldwide spread

Approach

In order to simulate how the Ebola epidemic might spread across the world, we have assembled a worldwide network based on international trade. Because the trade numbers are in US dollars, we have assumed a linear relationship between exports in dollars and travelers going abroad. This is a strong assumption, and we are considering more sophisticated mappings. We are using this as a long-range network representing trade-based population movement and connecting localized subnetworks with these trade-driven edges. We are developing a simulation framework to run in discrete time steps and predict the spread of Ebola across this worldwide network.

Trade network data preparation

The trade network dataset we are using is based on data retrieved from the web service API to the UN Comtrade database (**?**) via their web service API. The data queried for were country-to-country SITC-1 exports from the latest data available for each country.

Several problems with this raw dataset immediately became apparent which required working around. Many war-torn countries such as Liberia have not reported detailed export data to the UN in decades (Liberia last reported in 1984). As a result of this, the export totals are incorrect for the present day. In addition, this old data reports exports to countries that no longer exist, including East Germany and Yugoslavia. In order to make the dataset usable for modern-day predictions, we performed the following transformations on the data (using Perl scripts):

- Manually remapped non-existent countries to their modern equivalents. In a few cases, we had to do a best-effort mapping. For example, Yugoslavia split into many states, so we assign exports to historical Yugoslavia to modern-day Slovakia, since Slovakia currently has the largest economy of the states that once comprised Yugoslavia.
- Removed exports from states that no longer exist, preferring export data from modern equivalents.

We then needed to make the per-country exports sum up to the latest available export data. The United States **?** World Factbook contains up-to-date export totals for most countries in the world. Using the above data set, on the (admittedly strong) assumption that the export distribution from country-to-country in the UN Comtrade data set has remained the same between the last year a country has reported detailed export data and the latest totals from the CIA, we linearly renormalized each outgoing edge in our data set so that the total sum of exports equalled the latest CIA data for each country. This required a manual step of mapping country names in the UN dataset to those used in the CIA dataset.

Once we had a "renormalized" dataset including detailed export edges and totals matching the latest available data, we needed to map these numbers to theoretical international travelers. We could not find complete, or even nearly complete, publicly available data on the number of international outgoing travelers per country. While inbound tourism numbers are available from **?**, for the outbound tourism numbers from **?**, many countries are missing, especially the African nations that we care so much about from an Ebola outbreak perspective. Some data are available from the **?**, however they appear to be behind a paywall. Data for total air travellers are available from **?**, however this total includes domestic flights and so is less useful for our purposes.

Our current approach to map outbound dollars to outbound travelers is to assume a linear relationship between these numbers, and also to assume that the same linear relationship holds for imports to inbound tourists to a country. We currently use the United States as a model and use the ratio of imports to the United States per year to the number of tourists visiting the United States per year. Based on data from the **?**, 69.77 million people visited the United States in 2013. According to our renormalized data set, total imports into the United States were \$2.21 trillion during the same period. Dividing imports by visitors (an admittedly simplistic approach) gives us a scaling factor of approx. $31,665$. Therefore we have applied this scaling factor to all edges in our international exports network, giving us some approximation of outbound travellers from country to country based on export numbers.

Supplementing the trade network

Using a trade network has its flaws, even discounting errors stemming from poor approximations. For one, a trade-centric network ignores activity with low economic impact, such as a vegetable farmer traveling to a nearby country to sell his products at the market, or someone driving across the border to visit a family member. Relative to the economic impact of the industrial diamond or rubber trade (exports of Liberia) for example, these potentially disease-spreading behaviors simply are not represented equitably if at all in the model. On the other hand, economic trade data is widely accessible and is likely to be fairly accurate, and certainly it is safe to assume that that trade activity correlates with travel between two countries.

Due to the above drawbacks, we plan to supplement our trade network with a generated network that captures local commute-related movement as well. One approach to generating such a network is the gravity law, used for example to study traffic on Korean highways by **?**. The gravity law claims that the number of trips from point A to point B is inversely proportional to the square of the distance between them.

Simulation model

We are currently developing a simulation framework to incorporate intra-country spreading behavior (using compartmental and network-based models) with inter-country links based on the above worldwide network models. A related but distinct approach is the GLEaM stochastic simulation model developed by **?**. Our simulation model is a stochastic, discrete time-step model in which each time step is one day, and for each day, each infectious individual may travel to another country with some probability related to the number of people leaving the country via the worldwide networks on that day and the total population of the country. Within each country, infectious individuals spread their disease with some probability related to the intra-country model being used (compartmental or network-based).

APPENDIX

TABLE I: Parameter estimation for Ebola SEIR model (Guinea & Sierra Leone)

	Guinea			**Sierra Leone**		
Incidence Dependent Parameters	*Value*		*Comments*	*Value*		*Comments*
Initial Case t_0	December 2, 2013		one person fell ill	April 23, 2014		one person fell ill
S_0	0		-	0		-
E_0	0		-	0		-
I_0	1		-	1		-
R_0	0		-	0		-
C_0	1		-	1		-
Intervention time	March 2, 2014		Gov. of Guinea informed WHO	June 12, 2014		Country declared emergency
τ	110		-	50		-
Estimated Parameters	*Value*	*95% CI*	*Comments*	*Value*	*95% CI*	*Comments*
Incubation Time $1/k$	6.3	-	based on previous works	6.3	-	based on previous works
Infection Time $1/\gamma$	5.4957	[5.43, 5.545]	-	6.36	[6.324, 6.396]	-
β_0	0.2407	[0.2374, 0.244]	-	0.357	[0.3525, 0.3613]	-
β_1	0.2084	[0.2033, 0.2135]	-	0.2012	[0.1994, 0.2028]	-
q	32	[0.1, 100]	-	34	[1.9, 110]	-
Fatality Rate	0.67	-	-	0.38	-	-
R_0	1.323	[1.295, 1.341]	-	2.27	[2.243, 2.298]	-
R_1	1.145	-	-	1.28	-	-

TABLE II: Parameter estimation for Ebola SEIR model (Liberia & West Africa overall)

	Liberia			**West Africa**		
Incidenced Dependent Parameters	*Value*		*Comments*	*Value*		*Comments*
Initial Case t_0	March 31, 2014		official confirmation two infected	December 2, 2013		one person fell ill in Guinea
S_0	0		-	0		-
E_0	0		-	0		-
I_0	2		-	1		-
R_0	0		-	0		-
C_0	2		-	1		-
Intervention time	July 30, 2014		School shutdown	March 2, 2014		Gov. of Guinea informed WHO
τ	120		-	110		-
Estimated Parameters	*Value*	*95% CI*	*Comments*	*Value*	*95% CI*	*Comments*
Incubation Time $1/k$	6.3	-	based on previous works	6.3	-	based on previous works
Infection Time $1/\gamma$	10.7	[8.9,10.7]	-	6.8	-	-
β_0	0.169	[0.168, 0.191]	-	0.2	-	-
β_1	0.0001	[0.0001, 0.1]	-	0	-	-
q	0.0085	[0.0072-0.0264]	-	0	-	-
Fatality Rate	0.67	-	-		-	-
R_0	1.808	-	-	1.36	-	-
R_1	0.001	-	-	-	-	-