Predicting 2014 Ebola Outbreak in West Africa using Network Analysis

Shafi Bashar, Mike Percy, Romit Singhai

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

Abstract

The current Ebola outbreak in West Africa is the worst in history, and shows no signs of abating. At the time of writing, the number of cases in Liberia continues to grow exponentially, while linear growth continues in Sierra Leone.

Computer models allow for predicting outbreaks and can help agencies like the World Health Organization (WHO) allocate the resources and interventions needed to stem an outbreak. Most traditional epidemiological models are compartmental models which calculate the effective reproductive rate of an outbreak. We discuss several models, including the traditional SIR (Susceptible, Infectious, Recovered) model, and an SEIR model, where an additional Exposed state modeling a non-infectious incubation period is added to SIR. An SEIHFR model adds additional infective states related to transmission in a Hospital (H) and transmission due to a traditional funeral (F), modeling infection vector heterogeneity.

In addition to the compartmental models, we discuss a network model which avoids the random-mixing assumption employed by the compartmental models above. This is done by assigning each individual a finite set of permanent contacts. Another model that divides the world into geographical regions and models traffic flow due to transportation infrastructure is also discussed. Finally, we describe a contact network model for SARS that compares urban network, random network, and scale-free network models.

We organize the rest of the paper as follows. In section I, we provide a survey of previous relevant work in the area of epidemic prediction as summarized above including the pros and cons. In Section II we provide the project proposal and the planned direction of our research.

I. Literature survey

A. *The epidemiological SIR model (Kermack and McKendrick, 1932)*

The basis of the majority of research in epidemiological theory is based on the compartmental model. In a compartmental model, 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 compartmental model is the SIR model introduced in (Kermack and McKendrick, 1932). In the SIR model, three compartments or disease states for an individual are used:

- **S (Susceptible):** These are individuals who have not yet caught the disease. They are susceptible to infection following a contact with infectious individuals.
- **I (Infectious):** These are individuals who have the disease. They are infectious and have some probability of infecting each of their susceptible neighbors.
- **R (Recovered):** These are individuals who have experienced the full infectious period. These nodes are removed from consideration, since they are no longer infectious and are also considered to be immune from re-infection.

The changes among these states over time are represented by a set of differential equations. In order to capture the dynamics of disease spread over time, a population-wide random mixing model is assumed. In a random mixing model, the full population mixes at random, so that each individual has a small and equal chance of coming into contact with any other individual. 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.

Discussion. The SIR model is classic and effective, however is missing important features of the Ebola virus: local spreading, a non-infectious incubation period, and representation of different methods of transmission.

B. *The basic reproductive number of Ebola: SEIR model (Chowell et al., 2004)*

In (Chowell et al., 2004), the authors model the effect of Ebola outbreaks in Congo 1995 and Uganda 2000 using a compartmental model similar to the SIR model in Section I-A. 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 (Chowell et al., 2004) the basic SIR model is modified by adding an additional compartment "Exposed". The modified SIR model, i.e. the SEIR model presented in (Chowell et al., 2004) is reproduced in Figure 1.

In the SEIR model, susceptible (S) individuals in contact with the virus enter the exposed (E) state at a rate of $\beta I/N$. Here,

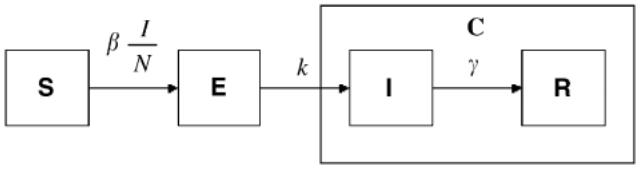

Fig. 1: SEIR model

β = transmission rate per person per day
N = total effective population size
$\frac{I}{N}$ = probability that contact is made with an infectious individual, assuming uniform random mixing

The exposed (E) individuals undergo an average incubation period of $1/k$ days before progressing to the infectious (I) state. The exposed state is assumed to be asymptomatic as well as uninfectious. Infectious (I) individuals move to the R state, either recovered or dead, at a rate of γ.

The following set of differential equations are used to represent this model:

$$\begin{aligned} \frac{dS}{dt} &= \frac{-\beta SI}{N} \\ \frac{dE}{dt} &= \frac{\beta SI}{N} - kE \\ \frac{dI}{dt} &= kE - \gamma I \\ \frac{dR}{dt} &= \gamma I \\ C &= kE \end{aligned} \tag{1}$$

Here, S, E, I, and R denote the number of susceptible, exposed, infectious and removed individuals at time t. In the equations, to simplify our notation, we have omitted the dependency on t. C is not an epidemiological state, however is useful to keep track of the cumulative number of cases from the time of the onset of the outbreak.

In order to model the effect of intervention on the spread of the disease, in the above model, the transmission rate β is modeled as a function of time. At the initial phase of the outbreak, before intervention, β is parameterized by β_0. After intervention, the value of β transitions from β_0 to β_1, $\beta_0 > \beta_1$ as follows:

$$\beta(t) = \begin{cases} \beta_0 & t < \tau \\ \beta_1 + (\beta_0 - \beta_1) \exp\left(-q\,(t-\tau)\right) & t \geq \tau \end{cases}$$

Where τ is the time when interventions begin and q controls how quickly the rate of transmission changes from β_0 to β_1.

The Ebola data for the 1995 Congo and 2000 Uganda outbreaks were represented as (t_i, y_i), $i = 1, 2, \ldots, n$ where t_i represents ith reporting time and y_i the cumulative number of infectious cases from the beginning of the outbreak of to time t_i. The model parameters $\Theta = (\beta_0, \beta_1, k, q, \gamma)$ were estimated using a least-square fit by fitting these data to the cumulative number of cases $C(t, \Theta)$ in Equation 1. The initial condition and appropriate of range of the parameters were taken from Empirical studies, e.g. an incubation period between 1 and 21 days and infectious period between 3.5 and 10.7 days were assumed. Once the parameters are estimated, the basic reproductive number was calculated using the following formula:

$$R_0 = \frac{\beta_0}{\gamma} \tag{2}$$

In addition to calculating R_0, (Chowell et al., 2004) also proposed an analogous continuous time Markov chain model based on the estimated parameters. The transition rates were defined as follows:

Event	Effect	Transition Rate
Exposure	(S,E,I,R)→(S-1,E+1,I,R)	$\beta SI/N$
Infection	(S,E,I,R)→(S,E-1,I+1,R)	kE
Removal	(S,E,I,R)→(S,E,I-1,R+1)	γI

The event times $0 < T_1 < T_2 < \dots$ at which an individual moves from one state to another are modeled as a renewal process with increments distributed exponentially,

$$P(T_k - T_{k-1} > t | T_j, j \leq k-1) = \exp(-t\mu(T_{k-1})) \tag{3}$$

Here,

$$\mu(T_{k-1}) = \frac{1}{\frac{\beta(T_{k-1})S(T_{k-1})I(T_{k-1})}{N} + kE(T_{k-1}) + \gamma I(T_{k-1})}$$

Based on the above stochastic model, (Chowell et al., 2004) provides the results of a Monte Carlo simulation that shows good agreement with the actual data.

Discussion. While the SEIR model adds needed features to the SIR model, it still suffers from the random mixing assumption and only represents a single infectious state.

C. *Understanding the dynamics of Ebola epidemics: SEIHFR model (Legrand et al., 2007)*

Similar to (Chowell et al., 2004), (Legrand et al., 2007) also studies the Ebola outbreaks in Congo in 1995 and Uganda in 2000. However, a major difference from (Chowell et al., 2004) is that (Legrand et al., 2007) modeled the spreading of disease in heterogeneous settings. In order to gain better insight of the epidemic dynamics, (Legrand et al., 2007) subdivided the infectious phase into three stages:

- Transmission of infection in community setting (I)
- Transmission of infection in hospital setting (H)
- Transmission of infection after death assuming a traditional funeral (F)

The SEIHFR compartmental model is reproduced in Figure 2, with transition rates following:

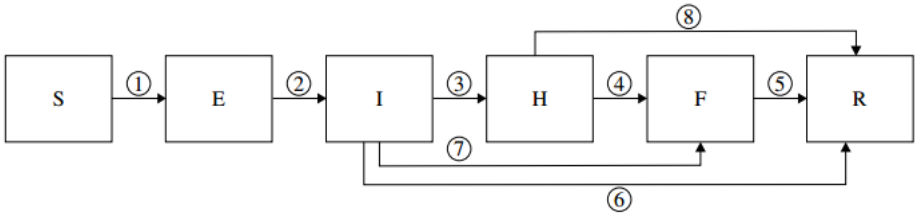

Fig. 2: SEIHFR model

Transition (i)	Effect	Transition Rate (λ_i)
1	(S,E)→(S-1,E+1)	$(\beta_I SI + \beta_H SH + \beta_F SF)/N$
2	(E,I)→(E-1,I+1)	αE
3	(I,H)→(I-1,H+1)	$\gamma_h \theta_1 I$
4	(H,F)→(H-1,F+1)	$\gamma_{dh} \delta_2 H$
5	(F,R)→(F-1,R+1)	$\gamma_f F$
6	(I,R)→(I-1,R+1)	$\gamma_i (1-\theta)(1-\delta_1) I$
7	(I,F)→(I-1,F+1)	$\delta_1 (1-\theta_1) \gamma_d I$
8	(H,R)→(H-1,R+1)	$\gamma_{ih} (1-\delta_2) H$

$\beta_I, \beta_H, \beta_F$ = Transmission coefficient in community, hospital and funeral respectively
θ_1 = fraction of infectious cases hospitalized
δ_1 = fatality ratio of infectious
δ_2 = fatality ratio of hospitalized patient
δ_1, δ_2 are computed such that overall fatality ratio is δ
$1/\alpha$ = mean duration of incubation period
$1/\gamma_h$ = mean duration from symptom onset to hospitalization
$1/\gamma_i$ = mean duration of infectious period for survivors
$1/\gamma_d$ = mean duration of infectious period to death
$1/\gamma_{dh}$ = mean duration from hospitalization to death
$1/\gamma_{ih}$ = mean duration from hospitalization to end of infectiousness for survivors
$1/\gamma_f$ = mean duration from death to burial

In order to model the effect of interventions, a two step approach is used:

- Before intervention, population was exposed to the cases in community, hospitalization as well as funeral
- After intervention, no transmission occurred at hospital or funeral, i.e. $\beta_H = \beta_F = 0$. The transmission coefficient in the community is decreased by a factor of $(1 - z)$.

In the above mentioned model, parameters $(\beta_I, \beta_H, \beta_F, z)$ were estimated by fitting the model to the morbidity data from the 1995 Congo and 2000 Uganda outbreaks using approximate maximum likelihood. The estimates of other parameters in the above model were drawn from prior work.

Simulations of the stochastic model were performed using Gillespie's first reaction method (Gillespie, 1976). At each iteration of the algorithm, a time τ_i is drawn from an exponential distribution with parameter λ_i for each of the transition. Here, λ_i is the transition rate of the transition i. The next transition μ is the transition that has the minimum time to occurrence (τ_μ). Counts in each compartment are updated accordingly. In addition to the simulation result, (Legrand et al., 2007) also presented the basic reproductive rate as a function of $(\beta_I, \beta_H, \beta_F, \gamma_h, \gamma_{dh}, \gamma_{ih}, \gamma_d, \theta_1, \delta_1, \delta_2)$.

Discussion. The SEIHFR model comes even closer to modeling the real-life behavior of Ebola, but due to being a compartmental model continues to suffer from the random mixing assumption. Additionally, as more states are added to the compartmental model, complexity increases and ease of characterizing and understanding the model is diminished.

D. Discussion: SIR (Chowell et al., 2004; Legrand et al., 2007) vs. networks (Newman, 2002)

(Chowell et al., 2004; Legrand et al., 2007), as described in previous sections, modeled the spread of Ebola using the compartmental modeling procedure. Even though (Legrand et al., 2007) modified the SEIR model to reflect the heterogeneity of infection states, the underlying assumption is still random mixing. A disease like Ebola spreads via networks formed by physical contact among individuals. While an individual may have the same number of contacts per unit time in either a random mixing model or a network contact model, within a static network model the set of contacts is fixed, versus a random-mixing model wherein it is continually changing. A static network model thus captures the permanence of many human relationships.

In (Newman, 2002), the authors extend the concept of the SIR model in network analysis. They provide an exact solution to the SIR model of epidemic disease on networks of various kinds. This is achieved using a combination of mapping to percolation models and using edge probability generating functions.

Transmissibility T of a disease is defined as the average probability that an infectious individual will transmit the disease to a susceptible individual with whom they have contact. The epidemic threshold T_c is the minimum transmissibility required for an outbreak to become a large-scale epidemic. The authors provided the relation between the basic reproductive number R_0 of an SIR network and the transmissibility T as follows:

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

In addition, (Newman, 2002) also provided the value of epidemic threshold T_c. In an uncorrelated network, it is given by:

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

Here, $\langle k\rangle$ and $\langle k^2\rangle$ are the mean degree and mean square degree of the network. Parameters for Poisson and power law networks are chosen such that all three networks share the same epidemic threshold. The authors also predicted average size of the outbreak $\langle s\rangle$ and probability of an epidemic S.

E. *Contact Networks for Epidemiology (Meyers et al., 2005)*

In our search for existing literature on the use of contact networks to model the spread of Ebola, we haven't come across any previous work that precisely does this, possibly due to the lack of detailed data in the locations historically affected by the disease. In (Meyers et al., 2005), the authors model the spread of the 2002-2003 outbreak of SARS in Hong Kong and Canada using a network model. A contact network model attempts to characterize every interpersonal contact that can potentially lead to disease transmission in the community. Each person in the community is represented as a node in the network and each contact between two people is represented as an edge connecting them. In (Meyers et al., 2005), the authors presented three different contact network models for SARS:

- **Urban Network:** A plausible contact network for an urban setting was generated using computer simulation based on the data for the city of Vancouver, British Columbia. $N = 1000$ households were chosen at random from the Vancouver household size distribution statistics which yields approximately 2600 people. Household members were given ages based on age distribution statistics and are then assigned to schools according to school and class size distributions, assigned to occupation according to employment data, to hospitals as patients and caregivers according to the hospital employment and bed data and to other public places.
- **Random Network:** The above urban networks offers high degree of realism, however is quite complex. In addition to the urban model, a random network with Poisson degree distribution in which individuals connect to others independently and uniformly at random.
- **Scale-free Network:** There may be individuals in the network called "superspreaders" with unusually large numbers of contacts or "supershedders" who are unusually effective at excreting the virus into the environment they share with others. Neither the urban nor the random networks contains significant number of superspreaders. To incorporate the effects of superspreader in disease transmission, (Meyers et al., 2005) also studies a network with truncated power law degree distribution. This type of network has a heavy tail of superspreaders. These individuals despite being few in numbers had a profound effect on the outbreak patterns.

Discussion. (Meyers et al., 2005) extended the result from (Newman, 2002) to calculate the fate of an outbreak based on its initial condition, the probability that a patient zero with degree k will start an epidemic, and the probability that an outbreak of size N will start an epidemic. However, (Newman, 2002) did not capture the temporal progression of the epidemic, instead providing an overall number and distribution of the infected individuals. The authors predicted the probability S that an outbreak with $R_0 > 1$ will lead to an epidemic for their three networks. S is often significantly less than 1 and can be different for two networks with the same R_0. Outbreaks are consistently less likely to reach epidemic proportions in power law networks than in the others. R_0 is a valuable epidemiological quantity, however it has its limit since R_0 is a function of both the transmissibility of a disease and the contact patterns that underlie the transmission. Therefore, measuring R_0 in a location where contact rates are unusually high will lead to an estimate that is not appropriate for the larger community. Estimating T instead of R_0 may give us a way out of this difficulty.

F. *Report assessing the international spreading risk of Ebola (Gomes et al., 2014)*

A recent study (Gomes et al., 2014) gathered data from the Disease Outbreaks News of the WHO and used what they call the Global Epidemic and Mobility Model to predict the spread of the current (2014) Ebola outbreak. In this model, the world is divided into geographcal regions defining a subpopulation network, modeling connections among subpopulations representing traffic flows due to transportation infrastructure.

Like (Legrand et al., 2007), they use the SEIHFR compartmental disease model and compare it with the more traditional SEIR model, while simulating an ensemble of possible epidemic evolutions for observables such as newly-generated cases, time of arrival of infection, and the number of traveling disease carriers. The parameter θ_1 is computed so that $\theta\%$ of infectious cases are hospitalized.

The expression for the basic reproductive number R_0 is obtained using the sum of three terms for this model:

$$R_0 = R_I + R_H + R_F$$

Where R_I is a term that accounts for transmissions in the community, R_H accounts for transmissions in the hospital, and R_F accounts for infections due to dead individuals. Parameters were fit using latin hypercube sampling of the parameter space defined by the vector $P = (R_I, R_H, R_F)$.

To make predictions, the authors ran Monte Carlo simulations exploring the value of R_0 while relying on results reported in (Legrand et al., 2007) and elsewhere for the rest of the model parameters.

Discussion. The approach used in (Legrand et al., 2007) appears to be the state of the art in epidemic prediction for situations where the available data is very limited. Unfortunately, as previously discussed, the SEIHFR model has a random mixing assumption that does not reflect reality over large areas. It may be possible to use small world networks or other generated networks to model locality more effectively based on population density and geographical information.

II. Project Proposal

A. Data set

The data set we are planning to use (Rivers, 2014) is an aggregation of Ebola outbreak data from multiple sources, including the WHO, the Liberia Ministry of Health, and the Sierra Leone Ministry of Health. The data set consists of time series totals for suspected infected, confirmed infected, and dead on an almost-daily basis across several countries, and counties within countries. To model a world-wide contact network, we will be using information from additional sources, such as the CIA factbook or World Trade Organization (WTO) economic trade data between countries.

B. Project plan

1) **Estimating model parameters and basic reproduction number for 2014 Ebola data using a random mixing model:** In the first phase of the project, we are planning to use the SEIR model proposed in (Chowell et al., 2004) as a baseline model, and estimate the model parameters $\beta_0, \beta_1, k, q, \gamma$ and basic reproduction number R_0 to fit the data obtained from current Ebola outbreak. Once the parameters are obtained, we are planning to use these parameters to perform prediction of the epidemic assuming a fully mixed model. To benchmark our prediction, we are planning to use the prediction performed by the CDC (Meltzer et al., 2014). We are interested in observing how different parameters in the model affect the spread of the disease. To this end, we are planning to vary critical parameter values, such as mean incubation period and transmission rate, in order to observe their effects on disease growth.
2) **From random mixing model to contact network:** (Meyers et al., 2005; Newman, 2002) extended the basic SIR model to network analysis. Whether this model is readily applicable to the modified (SEIR, SEIHFR) compartmental models used previously in Ebola research (Chowell et al., 2004; Legrand et al., 2007) is not immediately clear. In the second phase of our project, we will model contact networks using the basic reproductive numbers and other parameters obtained in phase 1. We will then estimate the propagation of disease on these contact networks by assuming an initial number of infected nodes. Since, in a contact network based model, each node can propagate the disease to only its immediate contacts, i.e. the neighboring nodes, the spreading of disease in a contact network based model will be different from our first phase of analysis, in which a random mixing paradigm is assumed.

   In the case of the currently-affected African countries, such as Liberia, we have county-level data points, and are planning to use that county-level data in different ways. One approach is to model individuals as nodes in a small-world network, and since the individuals within a particular county are more likely to have contact with each other, we may use a different probability of adding long-range links within a county than between counties. Another approach is to model each county as a node in a wide-area "internetwork". In any case, the number of edges between different counties can be dependent on different demographic parameters such as the underlying transportation networks, distance between counties, or amount of trade between the regions. Within each county, we also plan to simulate disease-spread behavior using different types of generated contact networks, including random graphs, small world networks, and scale-free networks.

   Thus far, the growth curve of the Ebola virus in Liberia has been exponential, while in other places like Guinea and Sierra Leone the growth curve is more linear (Rivers, 2014). When using a compartmental model for analysis, we will model different values of R_0. When modeling as a contact network, we can start by calculating a single value of transmissibility T and model contact networks for Liberia or Guinea differently using different average degrees of nodes. We can also try using different types of network models for different places and observe the effect of these different assumptions on the spread of the disease.

3) **Possibility of Ebola becoming a world-wide epidemic:** In (Gomes et al., 2014), the authors use airport network data to predict the spread of Ebola to different parts of the world. In the third phase of our project, we are planning to use a similar approach to predict the spread of the epidemic over a wider region of the world. To this end, we will make use of the estimation parameters and insight gained in phase 1 and phase 2 of our project. In order for the disease to spread, we need to consider the underlying contact networks between different regions of the world. In order to model such a contact network, we can make use of public data such as transportation networks, trade routes, and levels of trade between countries and regions. We may even be able to use demographic information. Since traditional funeral practices in certain regions may increase the chance of spreading the diease, such as setting the dead afloat in a river, we can also make use of river flow data.

References

Gerardo Chowell, Nick W Hengartner, Carlos Castillo-Chavez, Paul W Fenimore, and JM Hyman. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *Journal of Theoretical Biology*, 229(1):119–126, 2004.

Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

MF Gomes, AP Piontti, Luca Rossi, Dennis Chao, Ira Longini, M Elizabeth Halloran, and Alessandro Vespignani. Assessing the international spreading risk associated with the 2014 West African Ebola outbreak. *PLoS Currents Outbreaks*, 2014.

William O Kermack and Anderson G McKendrick. Contributions to the mathematical theory of epidemics. II. The problem of endemicity. *Proceedings of the Royal society of London. Series A*, 138(834):55–83, 1932.

J Legrand, RF Grais, PY Boelle, AJ Valleron, and A Flahault. Understanding the dynamics of Ebola epidemics. *Epidemiology and infection*, 135(04):610–621, 2007.

Martin I Meltzer, Charisma Y Atkins, Scott Santibanez, Barbara Knust, Brett W Petersen, Elizabeth D Ervin, Stuart T Nichol, Inger K Damon, and Michael L Washington. Estimating the future number of cases in the Ebola epidemic - Liberia and Sierra Leone, 2014-2015. *Morb Mortal Wkly Rep*, 63:1–14, 2014.

Lauren Ancel Meyers, Babak Pourbohloul, Mark EJ Newman, Danuta M Skowronski, and Robert C Brunham. Network theory and SARS: predicting outbreak diversity. *Journal of theoretical biology*, 232(1):71–81, 2005.

Mark EJ Newman. The spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

Caitlin Rivers. Data for the 2014 Ebola outbeak in West Africa, 2014. URL https://github.com/cmrivers/ebola. [Online; accessed 15-October-2014].