Predicting 2014 Ebola Outbreak in West Africa using Network Analysis

Shafi Bashar, Mike Percy, Romit Singhai

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

I. Introduction

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 are incredibly helpful in curbing an outbreak. They can help agencies like WHO to predict the resources and interventions needed to stem the outbreak. Most of the outbreak models are compartmental models and are used for calculating and reducing the virus effective reproductive rate. We discussed many model including SIR model S(Susceptible), I(Infectious) and R(Recovered). Also discussed is SEIR model where additional state modeling the incubation period (applicable for Ebola) is added to the model. The SEIHFR model adds additional state related to transmission in Hospitals(H) and traditional burial (F) and assumes heterogeneous setting in comparison to the homogeneous setting for the previous two models.

Apart from the compartmental models we also discussed a network model which avoids random-mixing assumption for the population by assigning each individual a finite set of permanent contacts.Another model called the Global Epidemic and Mobility Model which divides the world into geographical regions and models traffic flow due to transportation infrastructure is also discussed. Finally we describe a model based on contact network using Urban Network, Random Network and Scale-free network used for modeling SARS outbreak.

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

II. Reaction Paper

A. *SIR model for Epidemiology [?]*

The basis of majority of research in epidemiological theory is the based on compartmental model. In a compartmental model, to model the progress of an epidemic in a large population, the individuals in the population compartmentalized according to the state of the disease. The most widely used compartmental model is the SIR model introduced in [**?**]. In SIR model, three compartments or disease state for an individual is used:

- **S (Susceptible) :** These are individuals before catching the disease. They are susceptible to infection following a contact with infectious individuals.
- **I (Infectious) :** These are individuals who have caught 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 no longer pose threat of future 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 random mixing model, the population mixes at random, so that each individual has a small and equal chance of coming in contact with any other individual. A 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.

*B. **Paper [?]***

In [**?**], the authors model the effect of Ebola outbreaks in Congo 1995 and Uganda 2000 using compartmental model similar to the SIR model in Section II-A. However, a distinct feature of Ebola disease is, individuals exposed to the virus who become infectious do so after a mean incubation period. In order to reflect this feature, in [**?**], the basic SIR model is modified by adding an additional compartment "Exposed". The modified SIR model, i.e. the SEIR model presented in [**?**] is reproduced in Figure 1.

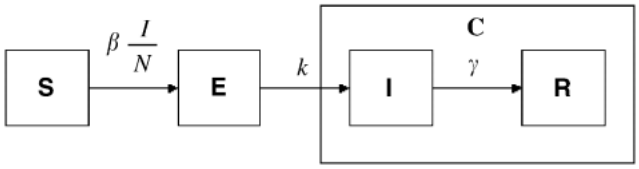

Fig. 1: SEIR model

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

$$\begin{aligned} \beta &= \text{transmission rate per person per day} \\ N &= \text{total effective population size} \\ \frac{I}{N} &= \text{probability that a contact is made with a infectious individual, assuming random uniform mixing} \end{aligned}$$

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) individual move to R state, i.e. either recovered or death 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 individual at time t. In the equations, for simplicity of notations, we removed the dependence of 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:

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

Here, τ is the time at which the interventions start and q control the rate of transmission from β_0 to β_1.

The Ebola data for Congo 1995 and Uganda 2000 outbreak 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 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 is calculated using the following formula

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

In addition of calculating R_0, [**?**] 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 $0 < T_1 < T_2 < \ldots$ 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, [**?**] provided Monte Carlo simulation, which shows good agreement with the actual data.

C. *Paper [?]*

Similar to [**?**], [**?**] also studies the Ebola outbreak in Congo 1995 and Uganda 2000. However, a major difference from [**?**] is that, [**?**] modeled the spreading of disease in heterogeneous settings. In order to gain better insight of the epidemic dynamics, [**?**] 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 during traditional burial (F)

The modified stochastic compartmental model is reproduced in Figure 2. The transition rate among different stages are provided in the following stochastic model:

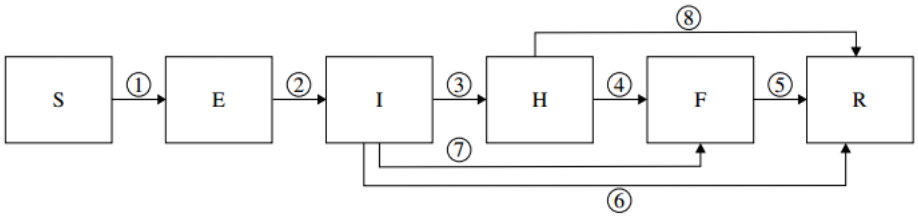

Fig. 2: SEIHFR model

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)	$\gamma_{ih}(1-\delta_2) H$

Here,

$$
\begin{aligned}
\beta_I, \beta_H, \beta_F &= \text{Transmission coefficient in community, hospital and funeral respectively} \\
\theta_1 &= \text{fraction of infectious cases hospitalized} \\
\delta_1 &= \text{fatality ratio of infectious} \\
\delta_2 &= \text{fatality ratio of hospitalized patient} \\
&\quad \delta_1, \delta_2 \text{ are computed such that overall fatality ratio is } \delta \\
1/\alpha &= \text{mean duration of incubation period} \\
1/\gamma_h &= \text{mean duration from symptom onset to hospitalization} \\
1/\gamma_i &= \text{mean duration of infectious period for survivors} \\
1/\gamma_d &= \text{mean duration of infectious period to death} \\
1/\gamma_{dh} &= \text{mean duration from hospitalization to death} \\
1/\gamma_{ih} &= \text{mean duration from hospitalization to end of infectiousness for survivors} \\
1/\gamma_f &= \text{mean duration from death to burial}
\end{aligned}
$$

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 of Congo 1995 and Uganda 2000 outbreak using approximate maximum likelihood. The estimates of other parameters in the above model were drawn from previous literatures.

Simulations of the stochastic model were performed using Gillespie's first reaction method [**?**]. 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, [**?**] also presented the basic reproductive rate as a function of $(\beta_I, \beta_H, \beta_F, \gamma_h, \gamma_{dh}, \gamma_{ih}, \gamma_d, \theta_1, \delta_1, \delta_2)$.

D. Discussion on [??] and Network Model for Disease Spread [?]

[**??**] as described in previous sections modeled the spread of Ebola using the compartmental modeling procedure. Even though [**?**] modified the original SIR model to reflect the heterogeneity of infection states, the underlying assumption is still the fully mixed model, where individual has an equal chance of spreading the disease to each other. However, disease like Ebola speeds through the populations via the networks formed by physical contacts among individuals. Models that incorporate network structure avoid the random-mixing assumption by assigning each individuals a finite set of permanent contacts to whom they can transmit the infection and from whom they can get infected. Although both in network and random-mixing based compartmental models, individuals may have the same number of contacts per unit time, within a network the set of contacts in fixed, whereas in random-mixing model, it is continually changing. A network model thus capture the permeance of interactions.

In [**?**], the authors extend the concept of SIR model in network analysis. They provide an exact solution to the SIR models of epidemic disease on networks of various kinds. This is achieved using a combination of mapping to percolation models and generating function methods.

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. 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, [**?**] 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. Assessing the international spreading risk associated with the 2014 West African Ebola outbreak [?]

Work by Gomes, et al. [**?**] 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 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 [**?**], they use the SEIHFR compartmental disease model and compare it with the more traditional SEIR model, while generating an ensemble of possible epidemic evolutions for observables such as newly generated cases, time of arrival of infection, and number of traveling disease carriers.

For the SEIHFR model, the model parameters are β_I, transmission coefficient in the community, β_H, transmission coefficient in the hospital, and β_F, transmission coefficient during funerals. θ_1 is computed so that $\theta\%$ of infectious cases are hospitalized.

The expression for the basic reprodctive 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.

They ran Monte Carlo simulations exploring the value of R_0 while relying on results reported in [**?**] and elsewhere for the rest of the model parameters. For the SEIHFR model, they used a latin hypercube sampling of the parameter space defined by the vector $P = (R_I, R_H, R_F)$.

F. Contact Network for Epidemiology [?]

In our search for existing literature on the use of contact network to model Ebola, we haven't come across any previous works. In [**?**], the authors model the spread of 2002-03 outbreak of SARS in Hong Kong and Canada using network model. A contact network model attempt to characterize every interpersonal contact that can potentially lead to disease transmission in community. Each person in community is represented as a node in the network and each contact between two people is represent as edge connecting the two nodes. The number of edges emanating from a node, that is the number of contacts a person has is called the degree of the node. The degree distribution of the node is a fundamental quantity in network theory. In [**?**], the authors presented three different contact network model for SARS outbreak:

- **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, [**?**] 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 profound effect on the outbreak patterns.

[**?**] extended the result from [**?**] to calculated the fate of an outbreak based on its initial condition, probability of a patient zero with degree k will start an epidemic and the probability that outbreak of size N will start an epidemic. However, [**?**] did not capture the temporal progression of the epidemic, rather provided overall number and distribution of infected individuals. The authors predicted the probability S that an outbreak with $R_0 > 1$ will lead to an epidemic for there three networks. S is often significantly less than one and can be different for two networks with the same R_0. Outbreaks are consistently less likely to reach and epidemic proportions in the 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 give us a way out of this difficulty.

III. Discussion Project Proposal

A. Proposal

1) **Estimating model parameters and basic reproduction number for 2014 Ebola data using [?] model**: In the first phase of the project, we are planning use the model proposed in [**?**] 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 CDC [**?**]. We are interested to observe how different parameters in the model affect the spread of the disease. To this end, we are planning to try out different configurations, e.g. the effect on disease spread on longer/shorter mean incubation period etc.
2) **From fully mixed model to contact network :** [**??**] extended the basic SIR model to network analysis. Whether this model is readily applicable to the modified compartmental model used previously in Ebola research [**??**] is not completely clear. In the second phase of our project, we will model contact networks using the basic reproduction numbers and other parameters obtained in phase one. We will then estimate the propagation of disease on these contact networks by assuming a initial number of infected nodes. Since, in contact network based model, each node can propagate the disease to only its immediate contact, i.e. the neighboring node, the spreading of disease in a contact network based model will be different from the first phase analysis where a fully mixed model will be assumed. We are planning to observe the behavior of disease spread in different types of contact networks, e.g. random network, small world network, scale-free network. The individuals within a particular county are more likely to have contacts with individuals in the same county. Therefore, we are planning to create a hierarchical networks among multiple counties. In this network, individuals in a county can be modeled as an small world network. The number of edges between different counties can be dependent different demographic parameters such as the underlying transportation networks, distance between counties etc.
3) **Modeling the Ebola data for network analysis :** So far, the spread of disease in Liberia is exponential whereas in other places like Guinea or Sierra Leon more linear. If we use compartmental model for analysis, then we will estimate different values of R_0. An alternative approach could be to calculate a single value of transmissibility T for network analysis and then model the contacts networks for Liberia or Guinea differently using different different average degree of nodes for each individual to fit the data. Alternatively, we can also try to use different types of network assumptions for different places and observe the effect of different assumptions on disease spread.
4) **Will Ebola become an world-wide epidemic? :** In **?**, 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 similar approach to predict the spread of epidemic over 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 contact network, we can make use of different available data, e.g. trade network, transportation network, demographic information, economic exchanges between country to country or region to regions etc. Since traditional burial may increase the chance of spreading, in addition, we can make use of river flow data.

B. Data Set

The data set we are using [**?**] 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 a nearly-daily basis across several countries and counties within countries. To model the world-wide contact network, we will be using

information from additional sources, e.g. CIA factbook or WTO economic trade data between countries etc.