

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. These models calculate 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 review a worldwide network model that represents traffic flowing across transportation networks and consider approaches for predicting the extent of an Ebola outbreak.

I. INTRODUCTION

TODO: Add introduction

II. RELATED WORK

Related work by (Gomes et al., 2014) 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 (Meyers et al., 2005), 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 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 (Kermack and McKendrick, 1932):

- **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 (Chowell et al., 2004), 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, the SIR model is extended with an additional “Exposed” compartment state. This SEIR model is summarized in Figure 1:



Fig. 1: SEIR model

In (Legrand et al., 2007), the Ebola outbreaks in Congo in 1995 and Uganda in 2000 are also studied. However, a major difference from (Chowell et al., 2004) is that (Legrand et al., 2007) models the spreading of disease in heterogeneous settings.

In order to gain better insight of 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).

This SEIHFR compartmental model is summarized in Figure 2.

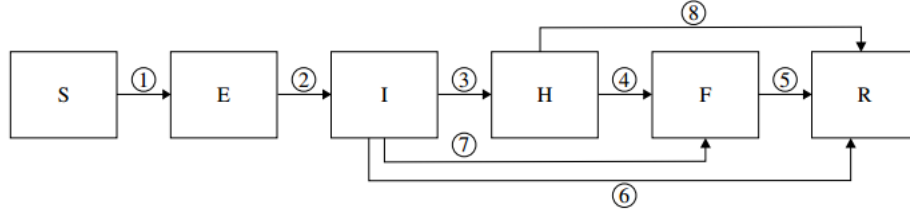


Fig. 2: SEIHFR model

III. MODELING INTRA-COUNTRY SPREADING BEHAVIOR

A. Compartmental approaches

TODO: We used SEIR and something

B. Network approaches

TODO: We are looking at modeling via G(n,m) and contact networks

IV. MODELING WORLDWIDE SPREADING BEHAVIOR

TODO: We are looking at modeling using trade data.

1) *Preparing data*: TODO: Used public data from the UN and CIA...

V. TODO: REMOVE BELOW DETAILS OR INCORPORATE INTO METHODS SECTION

A. SEIR model details

In the SEIR model, susceptible (S) individuals in contact with the virus enter the exposed (E) state at a rate of $\beta I/N$. 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 uninfected. Infectious (I) individuals move to the R state, either recovered or dead, at a rate of γ . The parameters are β , the transmission rate per person per day; N , the total effective population size; and I/N , the probability that contact is made with an infectious individual.

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

$$\frac{dS}{dt} = \frac{-\beta SI}{N}, \quad \frac{dE}{dt} = \frac{\beta SI}{N} - kE, \quad \frac{dI}{dt} = kE - \gamma I, \quad \frac{dR}{dt} = \gamma I, \quad C = kE$$

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 it 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:

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

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) , $i = 1, 2, \dots, n$ where t_i represents i th 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 ???. 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:

$$R_0 = \frac{\beta_0}{\gamma} \quad (1)$$

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	$(S,E,I,R) \rightarrow (S-1,E+1,I,R)$	$\beta SI/N$
Infection	$(S,E,I,R) \rightarrow (S,E-1,I+1,R)$	kE
Removal	$(S,E,I,R) \rightarrow (S,E,I-1,R+1)$	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,

$$P(T_k - T_{k-1} > t | T_j, j \leq k-1) = \exp(-t\mu(T_{k-1})) \quad (2)$$

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.

B. SEIHFR model details

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

Where the model parameters β_I , β_H , and β_F are the transmission coefficients in community, hospital and funeral settings respectively, θ_1 is the fraction of infectious cases hospitalized, and the rest of the parameters are omitted here for brevity.

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.

C. 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 provide 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 provide 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 predict average size of the outbreak $\langle s \rangle$ and the probability of an epidemic S .

The results in (Newman, 2002) allow us to compare more directly the relationships between random mixing models and network models under certain assumptions. This may prove handy in validating our models.

D. 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 urban-setting contact network was generated using simulation based on Vancouver data. Households were randomly chosen and their members given ages, schools, occupations, hospital beds as patients, and caregivers according to statistics from public data. This model offers a high degree of realism, but is complex.
- **Random Network:** A random network with Poisson degree distribution in which individuals connect to other 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 a significant number of superspreaders. To incorporate the effects of superspreaders 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 number, have 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.

E. 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 geographical 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 $R_0 = R_I + R_H + R_F$ 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.

VI. PROJECT PROGRESS

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 reproduction number of the current Ebola epidemic spread at different countries, we performed model fitting on the data we gathered from Rivers (2014). We considered the SEIR model described in Chowell et al. (2004). 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 anytime. We fitted our model to three countries in the West Africa - Guinea, Sierra Leon 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 the studies on previous Ebola epidemic. In Chowell et al. (2004), the incubation time of the Ebola $1/k$ is found to be varying between 1 to 21 days, with a mean time of 6.3 days for previous Ebola spread. 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 one, and therefore fixing a value based on the prior estimate may lead to some inaccuracy. To model the effect the 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 stage 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 fell in December 2, 2013, later diagnosed as Ebola patient. We consider this incidence as the index case for Guinea and set t_0 to December 2. In March 2, the Government of Guinea informed WHO regarding the possibility of Ebola epidemic and declared national health emergency. We considered this date as the intervention date and set τ to 110. In Sierra Leon, one person fell in April 2014. In June 12, 2014 the country declared emergency and closed borders with neighboring Guinea and Liberia. We consider the first date as t_0 and second date as intervention time, therefore set τ to 50. In Liberia, in March 31, 2014, there were official confirmation of two person getting infected from 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 in July 30, 2014. We consider this date as the date of intervention and set τ to 120.

In order to fit the non-linear SEIR model to data, we used non-linear least square estimation. We use the reported data (t_i, c_i) for $i = 1, 2, \dots, n$ where t_i denote the i -th reporting time and c_i as the cumulative number of infectious cases from the beginning of outbreak time t_0 to time t_i . The optimization problem contains a large number of local minimas. Therefore, the choice of initial parameter estimate is an important consideration to get the global optimum solution. In order to find a good initial choice of the parameters as an input to the non-linear least square 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 square error as the initial input. In order to calculate the 95% confidence interval of estimated parameter, we performed bootstrapping based on residual error.

The estimated parameters for the SEIR model of Guinea, Sierra Leon, Liberia, as well as West Africa region is presented in Table ?? . In Figure 3, we presented the the number of incidence 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 in October 21, 2014. As of today, November 11, 2014, few additional data points are available; we also plotted those data points in the graph. In addition to that, we extrapolated the graph up to December 31, 2014. We note that, the forecasting to future cases may not be appropriate 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 in par with the observed data for Guinea and Sierra Leon. We note that, for Guinea epidemic, we have data for the longest range of days as the initial case was in 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 under-estimated the observed data. Our guess is the estimated parameters for Libera 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 distorted the model fitting parameters. For the West Africa region plot, our prediction number is much higher that the actual data. This is understandable,

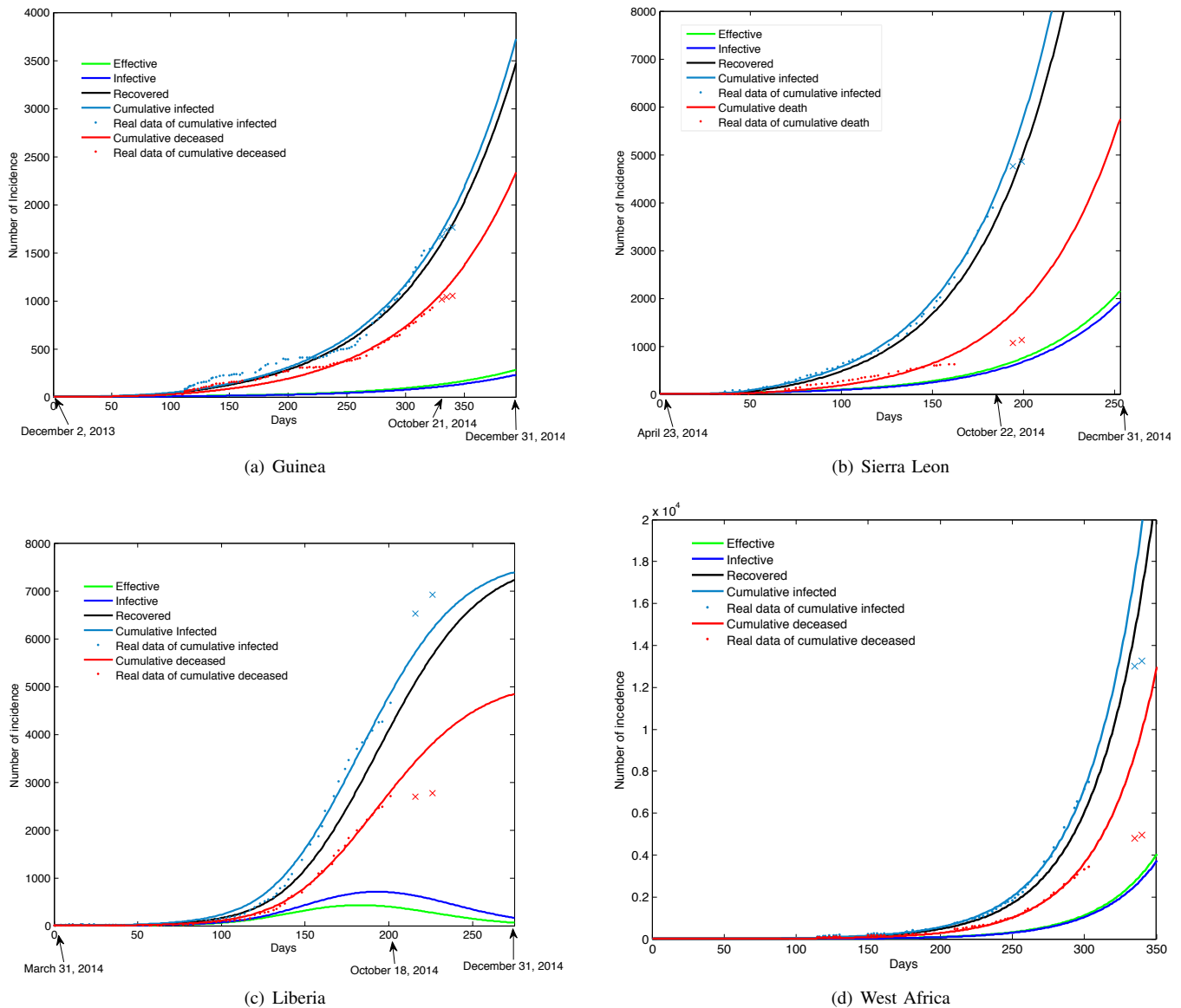


Fig. 3: SEIR model fit results for 2014 Ebola epidemic data

as the underlying assumption behind 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 random mixing assumption will over predict the infectious number of cases.

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.

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 outbreak in West Africa, 2014. URL <https://github.com/cmrrivers/ebola>. [Online; accessed 15-October-2014].

APPENDIX

TABLE I: Parameter estimation for Ebola SEIR model

Incidence Dependent Parameters	Guinea			Sierra Leon		
	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	-	-

Incidence Dependent Parameters	Liberia			West Africa		
	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	-	-	-	-	-