Modeling Ebola using a Macro- and Micro-level Network

Curran Kaushik (ckaushik) and Aaron Nagao (anagao)

CS224W Final Project Report

1 Introduction

The currently ongoing Ebola epidemic in West Africa is the deadliest and most persistent epidemic of Ebola in history. To study diseases like Ebola, researchers use theoretical models, and the standard is an SEIR compartmental model where people move through states of susceptible, exposed, infectious, and recovered. Assuming perfect mixing (a complete graph), the dynamics of these models can be represented by a series of differential equations, which are then used to predict the number of people in each state over time.

However, this complete-graph assumption fails to capture the relationships that are present in real-world networks, such as interactions between individual people or between the counties in a country. As such, SEIR models are able to forecast the number of cases over time, but cannot predict where Ebola will be most severe, or how severity might derive from a person or a county's connectedness within a network. To model this complexity, we propose a two-tiered geographic network that can model both macro-level interactions between counties as well as micro-level interactions within each county.

1.1 Problem Definition

Our goal is to determine if our original two-tiered network is a good theoretical model for disease spreading. We propose how a disease might spread across this two-tiered network (based on SIR and independent cascade), and then evaluate our proposed model by comparing a simulated spread of Ebola on our network to data from the current Ebola epidemic. After tweaking the parameters of our model, if we are able to model the Ebola epidemic accurately, then other researchers could use our two-tiered network to model the network effects inherent in other disease cascades.

1.2 Differences from Initial Proposal

We initially planned to infer a network structure from the Ebola data using the NETINF algorithm from Gomez-Rodriguez, Leskovec, and Krause (2012). However, NETINF required multiple datasets of disease cascades over the same region in West Africa as Ebola. Because detailed statistics on disease cascades are not curated well, especially in Africa, we decided to change our approach to defining a custom network structure and determining if it is accurate.

2 Related Work

2.1 Modeling the Current Ebola Epidemic

Since the Ebola epidemic is currently ongoing, any research on modeling it is extremely recent. Gomes et al. (2014) modeled Ebola using a SEIHFR model [2], which adds additional states to the basic SEIR model. They used prior literature on Ebola to set the parameters for their SEIHFR model and then ran a simulation on a global mobility network to predict how Ebola might spread internationally. Similar to Gomes et al., Rivers et al. (2014) also used an SEIHFR model, but focused on inferring the parameters for the model from the actual data rather than simply referring to past outbreaks [3]. They found the optimal model parameters that would best explain the data using weighted least-squares optimization, and then used their

optimal parameters to forecast how Ebola would spread according to their model. We will take a similar approach as Rivers et al., by using the existing data to formulate a more accurate model.

2.2 More Complex Models of Disease Spreading

In order to publish quickly, almost all predictions about the current ebola epidemic are based on SEIR models or reproductive number, which simplify the underlying network to be a complete graph or a tree. More complex models of disease spreading exist, and we will add our two-tiered approach to this literature. Lloyd et al. summarizes these other models [1], starting with "metapopulation models" that subdivide a well-mixed population into two subpopulations. These models typically still assume that the subpopulations are well-mixed, but also models the less-frequent mixing between the two subpopulations. This macro- and micro- layered approach forms the basis for our proposed network model, and we will add complexity by modeling the subpopulations with different models other than perfect mixing.

Lloyd et al. also contrasts models that focus on populations (like the metapopulation model) with models that focus on individuals, or network-based models. Network-based models model each member of the population as a node in a network, which requires significant complexity to fully describe not only each individual but also how all of these individuals interact with one another. By adopting network-based models at both our macro- and micro-layers, we encountered this complexity.

3 Approach

3.1 Ebola Dataset

Our Ebola dataset contained multiple CSV files and was obtained from the World Health Organization [4]. These CSV files contain the number of ebola cases reported each day from every county in Liberia, Guinea, and Sierra Leone. The number of days with available data varied by country, ranging from 58 to 139. We parsed and aggregated the data from various CSV files together, which enabled us to produce graphs showing the cumulative number of Ebola cases over time in each county. We used these graphs in our results section §4.4 to measure the success of our modeling, by comparing our simulation results to the actual data on the current ebola epidemic.

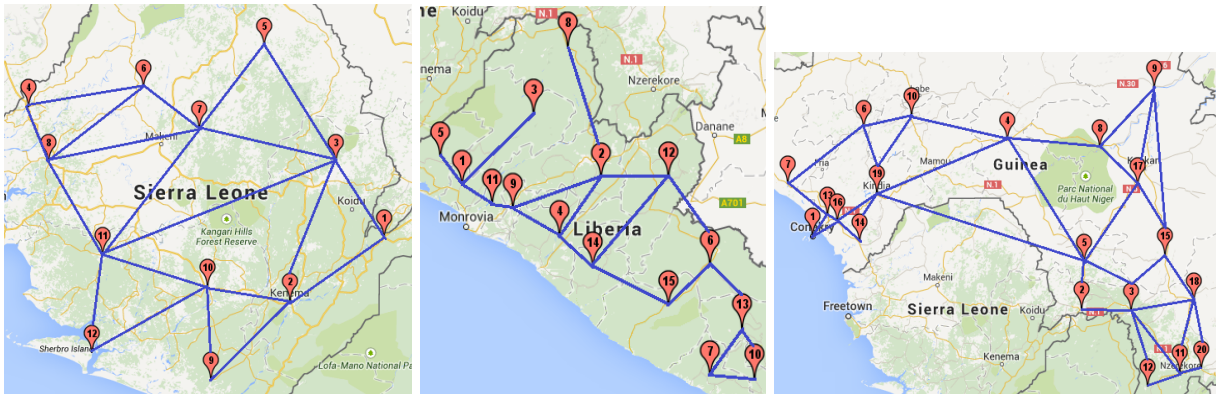
To put the absolute numbers of cases into perspective, we supplemented our dataset with the total population in every county from local census data [5]. As an illustrative example, by 11/06 there were 632 cumulative cases of Ebola in Bombali county of Sierra Leone, but this was a miniscule fraction of the 408,390 people in Bombali.

3.2 Network Modeling

To model the network on which Ebola will spread, we decided on a two-tiered network structure. The macro-level graph models relationships between counties, while micro-level graphs model relationships within each county between individual people. This two-tiered network allows us to model disease-spreading between counties and between people differently, which distinguishes us from previous research. Also, we note that each country (Sierra Leone, Liberia, and Guinea) was represented by its own independent two-tiered network.

3.2.1 Macro-Level Counties

On the macro-level, each node corresponds to a county, and edges represent a transportation connection between two counties. To build this network, we visually examined which counties were connected by major roadways on Google Maps, and these roadways became our edges. Figures 1-3 help provide a sense of how each county corresponds to a node and how the major roadways translated to edges between nodes.



Figures 1-3: Our macro-level graphs, where each node represents a county

We considered other models for our macro-level network, such as building a complete graph with edge weights indicating the distance between two counties, or similarly generating edges stochastically based on the distance between two counties. However, we ended up choosing our fixed approach partly due to simplicity, but also partly due to an assumption that the major roadways would be the primary routes along which Ebola would spread. Future research could investigate the effects of a different macro-level network representation.

3.2.2 Micro-Level Individuals

Every county was also modeled as a graph itself, with nodes representing people within a county and edges representing their connections. Figure 4 illustrates our general idea, where A,B,C are macro-level county nodes, and A1,A2,A3 represent the micro-level graph of people within county A:

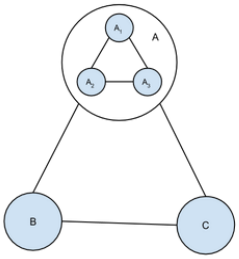


Figure 4: Our two-tiered network structure, with a micro-level graph (A1,A2,A3) embedded in each county of the macro-level graph (A,B,C).

More concretely, within every county, we have a node for every 10 people in the county (we allocated one node per 10 people for performance reasons that are discussed later). We chose to vary the edge structure within our micro-level graphs as an experimental parameter: we

experimented with modeling people within a county as a Watts-Strogatz small-world network, a complete graph, an Erdos-Renyi random graph, and a scale-free graph.

By modeling people within the county, we were able to measure the magnitude of an infection within a county (to a 10-person granularity), which could not be captured by the macro-level graph alone.

3.3 Simulation Algorithm

We simulated how Ebola propagates across our two-tiered network by using variants of independent cascade at the macro-level and SIR at the micro-level. Our algorithm follows:

1. In each country, start a cascade by randomly infecting an initial person-node in a random county
2. For each timestep:
    a. Micro-level (spreading between people within a county): SIR model
        i. For each Infected person-node, first determine if it moves from Infected to Recovered with probability δ. If it has not Recovered, then infect each of its neighbors with probability β.
    b. Macro-level (spreading between 2 counties): independent cascade model
        i. For each county, compute the proportion i of Infected people in the county, and then infect each of its neighbor counties with probability $q \bullet i$ (where q is some constant).
        ii. Infecting a neighbor county means picking a person-node at random within the county micro-graph, and if it is Susceptible, infect it.
    c. For evaluation, record the number of infected nodes in each county over time.

Because of our two-tiered network structure, our algorithm naturally fits into the MapReduce framework, where we can map over each micro-level county graph and simulate SIR within each county in parallel, and then a single reduce can simulate independent cascade among the various counties. This is another important feature of our model that allows for scalability. One could also imagine fitting this algorithm into some sort of graph-processing paradigm as supported by systems like Pregel/Giraph. Unfortunately, due to limited time and budget, we were unable to implement and deploy our simulation in a distributed fashion, and chose to leverage SNAP.py as our primary development tool instead. Our code is visible on Github at [5].

3.4 Evaluation

To evaluate our simulations, we compared the simulated number of cases in each county over each timestep with the actual number of cases. This evaluation was done both visually by comparing our reference graphs to the empirical plots generated by our simulation, and numerically by computing the root-mean-squared-error between our empirical number of cases and the reference number of cases, summed over every county and every day.

3.5 Grid Search over Parameters

To find the best set of parameters, we performed a grid search over all combinations of parameter values, where the grid was restricted to reasonable values for each parameter (where

reasonable is justified below). The set of parameters that minimized RMS error was thus our most accurate model for Ebola. Furthermore, because each country was represented by its own independent network, each country also had a different set of optimal parameters, as these different parameters modeled how the spread of Ebola varied in each country.

Parameter	Description	Grid search domain
δ	SIR probability a person-node transitions from Infected to Recovered	{0.6, 0.7, 0.8, 0.9}
β	SIR probability an infected person-node infects its neighbor person-nodes	{0.02, 0.04, β from Rivers et al. [3]}
q	Scaling factor for the proportion of infected people, to get the probability a county-node infects its neighbor county-nodes	{55, 150, 400, 2000}
micro-level graph model	Which theoretical model matches the network within a county/between people	{small-world, complete graph, random, scale-free}
k	Node degree of each person	{10, $R_0/\beta \approx 1.5/\beta$}

3.5.1 Justification for grid search domains

δ is the probability that a person-node recovers: in our SIR model, he cannot get re-infected and also no longer infects his neighbors. Due to the simplicity of our three-state SIR model, this parameter encompasses people who die from Ebola as well as people who recover and survive. The case fatality rate ranges from 0.48 in Sierra Leone to 0.74 in Guinea [6], so our δ must already be at least as large as this. We settled on {0.6, 0.7, 0.8, 0.9}.

Likewise, β is the probability that an infected person infects his neighbors. Rivers et al. [3] modeled this probability as a combination of contact rates from the community, from within a hospital, and from funeral/burial rites, spread across an incubation period. Her model learned that the probability of being exposed to Ebola is 0.319 for Sierra Leone and 0.711 for Liberia and we divide this by an incubation period of 10-12 days to get β=0.0319 in Sierra Leone and β=0.059 in Liberia. Thus we chose β around this range.

For the q scaling factor, the actual data showed that the proportion of infected people in a county was about 1/1000 (~200 infected people in a county of ~200,000). So we chose scaling factors that would give reasonable probabilities for inter-county spreading, ranging from 55/1000 = 5% to 2000/1000 = 100%.

For the node degree of the people within a county, the second lecture of 224W showed that most real-world networks have average degree around 10 [7]. We also tried another value derived from the reproductive number of Ebola $R_0 \approx 2$. If, on average, an Ebola patient should infect 2 other people, and at every timestep he infects each of his neighbors with probability β, then giving him 2/β neighbors would mean that at every timestep he is expected to infect 2 people, matching the reproductive number. However, this over-estimates the strength of Ebola because he infects 2 people at every timestep, so we scaled this down to give a person only 1.5/β neighbors.

4 Results

There were two high-level findings that we encountered early on that directly impacted the types of simulations we were able to run, and therefore the data that we were able to collect.

4.1 Scalability

The first observation was that modeling each individual as a single node would not be feasible in our SNAP.py implementation due to performance limitations (specifically, insufficient memory). Each county has hundreds of thousands of people, which means graphs with hundreds of thousands of nodes, and there would need to be almost 50 such graphs, one corresponding to each county. We tried to alleviate this problem by running our simulations on larger Amazon EC2 instances as opposed to our personal computers, but we were unable to find a solution that would allow us to achieve this level of granularity in the graph without spending large amounts of money on additional hardware resources. Thus, we settled on every micro-level node representing 10 people.

4.2 Varying micro-level network model

The second, more interesting observation that we quickly arrived at was that our micro-level networks must be small-world networks for our empirical results to even be in the same order of magnitude as the reference data. When a county was modeled as a complete graph, random network, or scale-free network, we observed that the county extremely rapidly transitioned from getting its first Ebola case to the entire county becoming infected with Ebola (this transition typically took only 3 to 4 timesteps). This phenomenon persisted even when we decreased the infection probability, increased the recovery probability, and decreased the probability of inter-county transmission within reason.

Network theory from CS224W can explain why the complete graph, random, and scale-free models all suffered from such rapid and complete infection. Both the complete graph and random graph have smaller diameters than the corresponding small-world network: a complete graph has diameter 1, and the small-world network effectively "trades off" some of its low diameter for an increased clustering coefficient when compared to the random network. The scale-free network similarly has low diameter, but its observed rate of infection was even more severe (2 to 3 timesteps from initial to complete infection). This could be explained by the presence of high-degree hubs in scale-free networks. The probability of a node having an edge to some other node is proportional to that other node's degree: $P(i \rightarrow j) \propto d_j$. Thus, a hub is likely to be infected quickly (because it is exposed to a lot of people), and once it is infected, it is likely to infect a lot of people all across the network. Because these three classes of networks were overrun by Ebola extremely quickly, all data and findings presented below are from simulations where the counties are modeled as small-world networks.

In addition, by fixing our micro-level model within each county as Watts-Strogatz small-world networks, an additional parameter we could vary was the rewiring probability: when constructing a small-world graph, the probability that an edge from the initial ring lattice would be randomly rewired. Conceptually, smaller values of this probability would increase the diameter of the graph and cause ebola to be less widespread within a county. We varied this parameter among {0.010, 0.003, 0.001}, but ultimately found that the standard rewiring probability of 0.01 performed the best.

4.3 Set of Optimal Parameters

Parameter	Sierra Leone	Liberia	Guinea
δ (infected → recovered)	0.7	0.8	0.6
β (infect people)	0.02	0.02	0.02
q (infect cross-county)	400	55	150
micro-level graph model	small world	small world	small world
k (node degree)	10	10	10

Our final root-mean-squared-error values for each country respectively were 548, 403, and 161.

4.4 Final Results

For each of our simulations, we produced graphs depicting the cumulative number of Ebola cases over time in all counties. The resulting graphs are displayed below. For each of the three countries, we have included the reference graph for that country (as described in §3.1), alongside the graph produced by our best simulation for that country (as defined by the optimal parameters in §4.3).

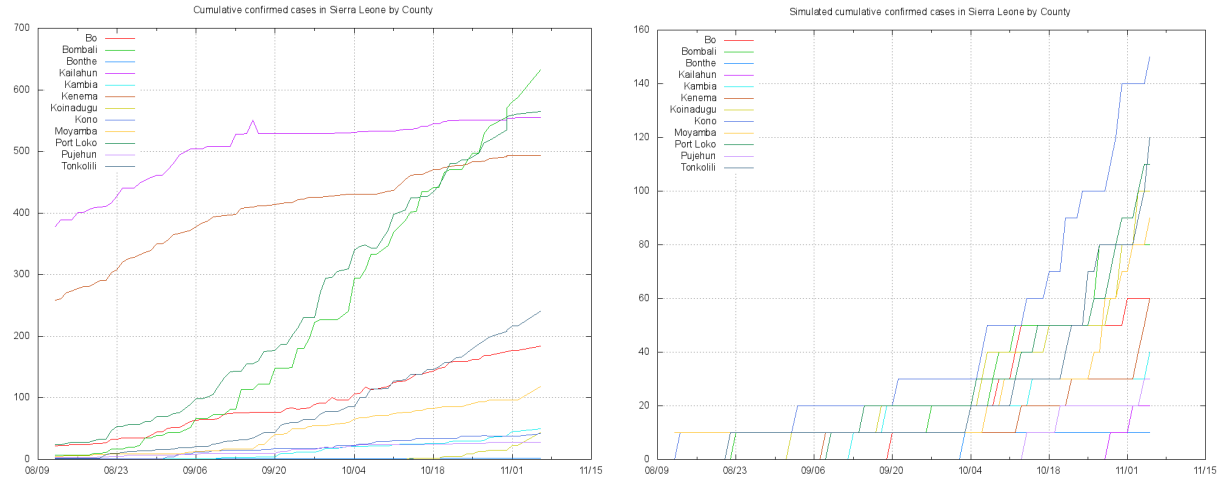


Figure 5: For Sierra Leone: the actual (left) and simulated (right) number of Ebola cases over time

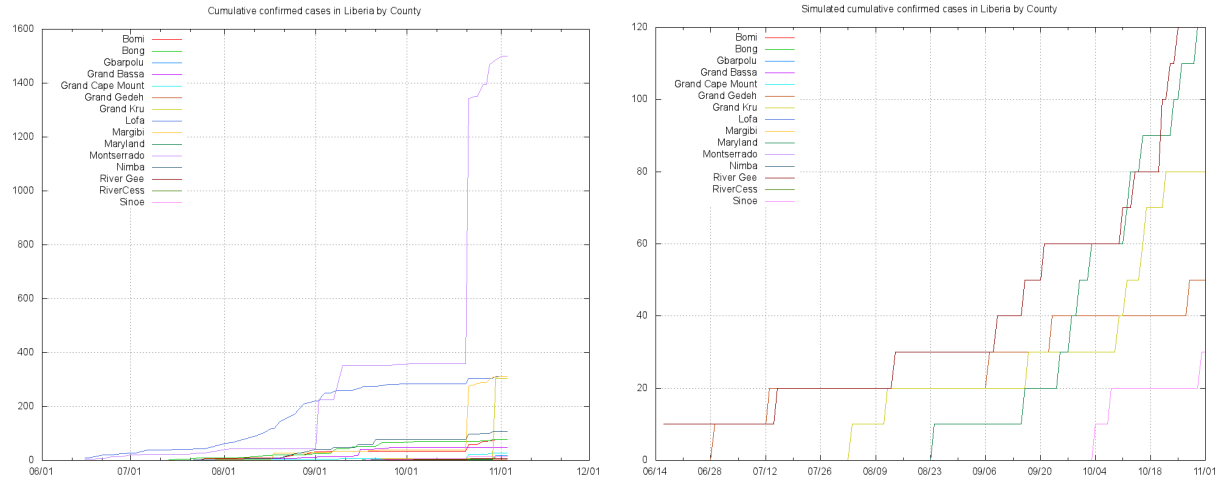


Figure 6: For Liberia: the actual (left) and simulated (right) number of Ebola cases over time

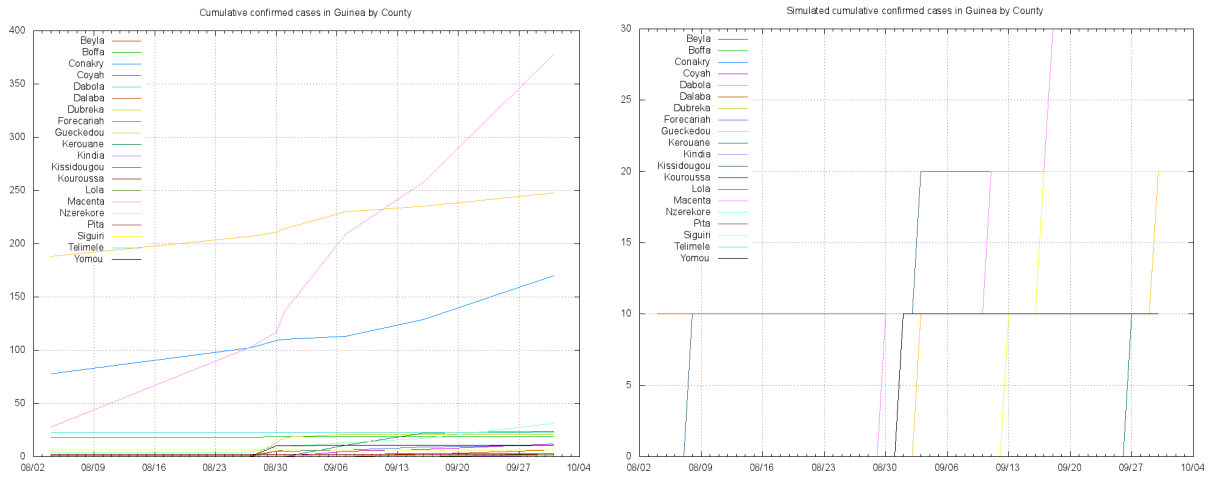


Figure 7: For Guinea: the actual (left) and simulated (right) number of Ebola cases over time

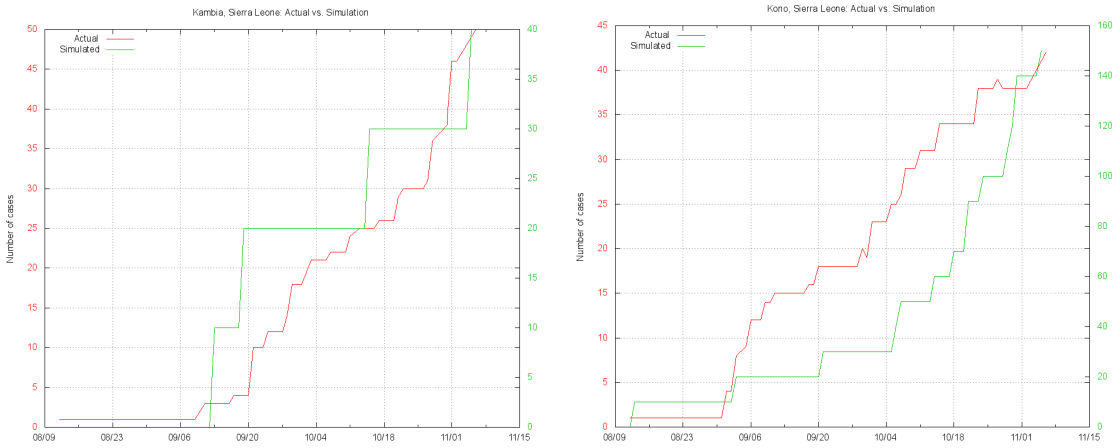
Because our model was not extremely accurate, we were less concerned with overfitting. Given that we were trying to match a very large set of datapoints (the number of Ebola cases in each county on every day), and given that we had a simplified independent-cascade and SIR model, overfitting with our model was unlikely.

5 Analysis

5.1 How accurate was our micro-graph model?

As seen in the graphs above, our small-world networks did not accurately model the magnitude of Ebola cases in each county over time. Although only 1/1000 people in each county were actually infected with Ebola, as we mentioned earlier almost everyone in the county was infected in our first round of simulations, and we struggled to bring this number down while maintaining reasonable values of each parameter. In the end, we were able to predict infected cases in the same order of magnitude, but no more precision than that.

However, the small-world model predicted the general rates of infection reasonably well. The two graphs below show the actual vs. simulated results in two counties of Sierra Leone, and although the scales of the y-axes are slightly different, the curves follow the same general trend.



Figures 8 and 9: The actual (red) vs. simulated (green) results in Kambia and Kono, Sierra Leone.

The left graph of Kambia shows a slow start (influenced by our macro-graph structure), and then a roughly linear increase in the number of cases over time. Perhaps if we were not limited by our 10-person level of granularity, the graphs would have matched quite closely. From the right graph of Kono, the interplay between our macro- and micro-level networks was able to predict periods of stasis where there are no new infections, such as from 09/20 to 10/04, and periods of new infections that vary in number (some days have 5 new infections, followed by 15, followed by 10). Because the actual and simulated curves have the same general trends, this validated our choice of a small-world network to model interactions between people within a county.

5.2 How accurate was our macro-graph model?

Recall that the macro-graph is responsible for the transmission of Ebola across different counties. As a result, to evaluate the efficacy of the macro-graph structure, we need to examine how the number of Ebola cases in each county compare to each other. One qualitative approach to perform this evaluation is to rank counties by their cumulative number of cases at the end of the simulation, then compare these rankings to those from the reference data. The table below shows the three counties with the highest number of cumulative cases for both our simulation data (collected from the simulation for each country with lowest RMS error) and the reference data. Similarities between the simulations and reference data are displayed in green.

	Reference			Simulation		
	Sierra Leone	**Liberia**	**Guinea**	**Sierra Leone**	**Liberia**	**Guinea**
1	Tonkolili	Montserrado	Macenta	Kono	River Gee	Macenta
2	Port Loko	Lofa	Gueckedou	Tonkolili	RiverCess	Gueckedou
3	Kailahun	Grand Kru	Conakry	Port Loko	Grand Kru	Siguiri

As the above table indicates, our macro-graph structure was fairly successful in producing top-3 rankings that matched those from the reference data. In other words, our macro-graph structure was reasonably capable of replicating the actual trend of which counties suffered from the highest number of Ebola cases. Furthermore, when examining our output simulation data, we found that these results were not simply caused by our algorithm “getting lucky” and choosing to infect patient zero from one of these counties. Rather, it appears that the inter-county interactions over time, as defined by our macro-graph model, successfully yielded these rankings.

6 Conclusion

Our attempts to model the spread of Ebola within and across the counties of Sierra Leone, Liberia, and Guinea did not achieve our lofty goal of exactly modeling how the current Ebola epidemic has spread across West Africa. Nevertheless, our project still yielded lessons that could be applicable to future approaches of modeling Ebola and other diseases.

First, we learned that graph models with very low diameters are unsuitable for representing the connections between individuals in a county. Instead, models like the small-world network must be used to avoid simulations that forecast a rapid and complete infiltration of all nodes by a disease.

Although the micro-level networks did not match the magnitude of cases in the reference data, predicting either much too high or much too low, there were some encouraging instances where the trend in the rates of new Ebola cases matched the reference data. We note that it is possible that our micro-graph approach could have produced improved results with finer granularity nodes (i.e. 1 node per person). A parallelizable implementation of our simulation could allow for this increased granularity, and for future research we outlined how our simulation can fit into the MapReduce framework.

Similarly, our simulations indicate some promise in macro-level network modeling, as it was able to correctly identify counties that would have the most cases of Ebola. Even this information alone is quite meaningful, as one can apply it in a predictive manner to inform the optimal allocation of resources to combat an epidemic. In our simulations, we fixed a macro-level structure and focused primarily on varying properties of the micro-level networks, but in future work it would be worthwhile to examine other macro-level network structures, such as more dynamic, stochastic graphs. After all, networks are constantly evolving, and so should our models.

7 References

[1] A.L. Lloyd, S. Valeika, A. Cintron-Arias. “Infection Dynamics on Small-World Networks,” *Modeling the Dynamics of Human Diseases: Emerging Paradigms and Challenges*, 2006.
[2] M.F.C. Gomes, A.P. Piontti, L. Rossi, D. Chao, I. Longini, M.E. Halloran, A. Vespignani. “Assessing the International Spreading Risk Associated with the 2014 West African Ebola Outbreak,” *PLOS Current Outbreaks*, September 2014.
[3] C.M. Rivers, E.T. Lofgren, M. Marathe, S. Eubank, B.L. Lewis. “Modeling the Impact of Interventions on an Epidemic of Ebola in Sierra Leone and Liberia,” 2014.
[4] C.M. Rivers. Data for the 2014 ebola outbreak in West Africa. https://github.com/cmrivers/ebola
[5] Project Code Repository: https://github.com/currankaushik/cs224w_final_project
[6] C.L. Althaus. “Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa,” *PLOS Current Outbreaks*, September 2014.
[7] J. Leskovec, et al., Internet Mathematics, 2009.