

UVM ScholarWorks

Infectious Disease Exploration with Operation Outbreak Simulations

Item Type	undergraduate thesis
Authors	Blanchard, Nathan James
Download date	2026-05-19 20:04:29
Item License	http://creativecommons.org/licenses/by-nc-nd/3.0/
Link to Item	https://hdl.handle.net/20.500.14849/5644

Infectious Disease Exploration with Operation Outbreak Simulations

Nathan Blanchard

Undergraduate Honors Thesis

Research Mentor: Jean-Gabriel Young Ph.D.

Thesis Committee Member: Laurent Hébert-Dufresne Ph.D.

**College of Engineering & Mathematical Sciences
The University of Vermont**

April 30, 2025

Abstract

Operation Outbreak is a mobile app-based simulation tool that uses Bluetooth to transmit a virtual virus between individuals in close proximity. This offers access to simulated disease transmission data within a controlled environment, enabling analyses that would be unfeasible with data from real-world outbreaks. In this research, we assess the validity of Operation Outbreak as a tool for obtaining realistic transmission trees by making comparisons between Operation Outbreak simulated infection trees and transmission dynamics that have been observed in real-world infection trees. Our findings indicate that the analysis of Operation Outbreak simulations produces results consistent with real-world transmission dynamics, including superspreaders infecting other superspreaders more than expected, a decrease in R_0 over time during an outbreak, and a decrease in the proportion of cases causing superspreading events. We then briefly explore some of the unique opportunities for analysis enabled by Operation Outbreak simulations, which would be difficult to accurately replicate using real-world data. These include mixing pattern analysis based on individual protection levels, observing changes in serial intervals, and determining conditional probabilities of infection by looking at the entire simulated population. Our research enables the use of Operation Outbreak simulations as a tool to better our understanding of disease transmission dynamics and inform responses to real-world outbreaks.

Contents

1	Introduction	4
2	Operation Outbreak Dataset	5
2.1	What is the data?	5
2.2	How was the data prepared?	6
2.3	Network Analysis Tools	6
2.4	Infection Trees	6
3	Evaluating the Realism of Operation Outbreak Simulations	7
3.1	Parameter Estimation	7
3.2	Superspreader Analysis	10
3.3	First vs. Second Half of Tree	11
4	Overcoming Limitations Through Operation Outbreak	14
4.1	Mixing Pattern Analysis	14
4.2	Serial Intervals	15
4.3	Conditional Probability of Infection	17
5	Conclusion	18
A	Inclusion Criteria	22
A.1	Participant Inclusion Criteria	22
A.2	Simulation Inclusion Criteria	22
A.3	Final Simulation Selection	22
B	Data Table	22
C	Goodness of Fit for Negative Binomial Distribution	22

1. Introduction

Understanding and forecasting disease spread is crucial for effective public health responses. However, relying on real-world transmission data for building models comes with significant limitations. These transmission trees, which track the spread of disease between individuals, are often limited, small, and noisy, containing missing, incomplete, and inaccurate data. This poses challenges for accurate analysis and disease forecasting.

Operating through a mobile app, the Operation Outbreak simulation unleashes a virtual pathogen through Bluetooth to simulate how outbreaks spread between people [1]. Participants download the app, which uses Bluetooth to transmit the virtual virus to other users in close proximity, mimicking real-world disease transmission. Operation Outbreak addresses the limitations of real-world transmission trees, providing access to high-quality trees in a controlled setting. This allows for more accurate analysis and forecasting and provides access to transmission data that we could not dream of with a real-world infection tree.

This raises the central question of our research: Does Operation Outbreak simulation data realistically capture key transmission dynamics observed in real-world infection trees?

In analyzing Operation Outbreak simulations, we will be taking a network approach to epidemiology. Danon et al. [2] offer a comprehensive overview of the application of network theory to epidemiology of infectious diseases. The authors split this paper into four sections: relevant networks for epidemiology, characterizing these networks, the statistical methods that can be used with these networks to infer epidemiological parameters, and how to determine epidemic dynamics using simulation and analytical methods over a network. This shows how, through network epidemiology, it is possible to better understand the spread of disease and form more accurate models for predicting the spread of an infection. Additionally, network models, or graph models, naturally represent how individuals in a population interact with one another [3]. Another benefit that the use of network models in epidemiology presents, as explained by Hébert-Dufresne et al. [4], is that the network approach naturally accounts for heterogeneity. This paper talks about the importance of considering heterogeneity while assessing risk and creating forecasts for disease spread. The reproductive number, R_0 , is often used in isolation as the sole metric for these assessments and forecasting, though the distribution of secondary infections can explain why outbreaks with similar R_0 values can greatly differ in size. The authors explain that when considering the entire range of secondary infections caused by an infected person, there is no straightforward connection between R_0 and the magnitude of an outbreak, pushing for a network approach to epidemiology which naturally accounts for this heterogeneity.

Many of the methodologies that we plan to use during this research are inspired by related works. What has probably been most influential in our research is a study by Taube et al. [5] that details the development and analysis of an open-access database of infectious disease transmission trees called OutbreakTrees. This database consists of 382 transmission trees for 16 different directly transmitted diseases, each consisting of anywhere from 2 to 286 nodes. Their analysis using this dataset was performed on a subset of 39 larger trees. They observed changes in parameters R_0 (the average number of secondary transmissions per case) and k (the dispersion parameter detailing the variation in the number of secondary infections per case) over time and found a significant decrease in R_0 and observed fewer cases causing superspreader events between the first and second halves of the transmission trees included in their analysis. Additionally, they found that, in two-thirds of the trees included in their analysis, superspreaders infected other superspreaders at a higher-than-expected rate.

In a related study using data from Operation Outbreak, Specht et al. [6] show how we can use simulation technology such as Operation Outbreak to mitigate and preempt viral outbreaks. The simulations studied in this paper are two simulations from Colorado Mesa University (CMU) and Brigham Young University (BYU). These results show the importance of considering an individual's second-degree contacts in predicting their individual-level risk of infection. Although there is a fairly strong relationship between first-degree and second-degree contacts, some individuals are at a higher or lower risk than their number of first-degree contacts would suggest. Using this information, public health authorities could identify which individuals are most likely to become infected and prioritize the deployment of preventative measures to those individuals (such as masks or vaccines). However, this study does not examine changes in outbreak parameters over time or explore superspreader dynamics, nor does it benchmark results against real-world infection data. As a result, it does not address our specific research questions.

This research will analyze the disease simulations provided by Operation Outbreak, benchmarking the structural properties of infection trees from Operation Outbreak against real-world transmission networks like those explored by Taube et al. [5] in their study of OutbreakTrees. In Section 2, we will briefly explore how the Operation Outbreak data is formatted, as well as discuss how we selected and prepared infection trees for analysis. In Section 3, we will estimate common disease parameters such as the reproduction number R_0 , the dispersion parameter k of a negative binomial distribution. We will also explore disease transmission dynamics by looking at superspreaders and the proportion of cases causing these superspreader events, following the definition of a superspreader by Lloyd-Smith et al. [7], and assessing if superspreaders infect other superspreaders at a higher-than-expected rate. Then, to see how transmission dynamics and disease parameters change over time, the parameters from the first and second halves of these simulations will be compared.

Our goal is to evaluate whether Operation Outbreak simulations reflect real-world infection trees. If they do, these simulations could provide a valuable tool for studying transmission dynamics in ways that would be difficult or unfeasible to achieve with real-world outbreak data. In Section 4, upon finding that Operation Outbreak does successfully mimic real-world infection trees, we will discuss some of the limitations in studies such as OutbreakTrees that the use of Operation Outbreak simulations can address. Then, to demonstrate the potential of studying disease transmission dynamics using simulated infections such as Operation Outbreak, we will perform mixing pattern analysis with respect to individual protection level, analyze serial intervals for disease transmission, and attempt to compute conditional probabilities for infection given different levels of protection.

This research will contribute to the field of network epidemiology by exploring the possibilities enabled by having access to high-quality infection trees in a controlled setting, and by suggesting that Operation Outbreak may be a viable way to obtain such data. In much of the related literature, researchers have faced limitations that are present with using real-world infection trees. Danon et al. [2] acknowledge that analysis of real-world infection trees is hindered by the limited data collected on contact network dynamics and the inability to collect data on an entire population in many cases. In the analysis of OutbreakTrees carried out by Taube et al. [5], much of their research, including comparing the first and second halves of infection trees for example, was performed under the assumption that the trees were incomplete, and they acknowledged limitations such as how transmission trees are typically an incomplete part of a larger chain of transmission events and how information about how control measures and behavior changes alter disease parameters in the middle of an outbreak is not available. Based on our analysis of Operation Outbreak simulations, we suggest that these simulations replicate key structural properties of real-world infection trees, particularly regarding superspreading dynamics and temporal changes in transmission parameters. This suggests that simulated outbreaks could be used as a tool for studying disease spread and informing public health interventions, addressing many limitations faced when analyzing real-world infection trees.

2. Operation Outbreak Dataset

2.1. What is the data?

Our analysis primarily relies on the Operation Outbreak histories dataset, which records 72 simulations, each with events generated by participants throughout a simulation. This gives us timestamps of when users are infected, how they were infected, who they were infected by, and their protection status at time of infection. This allows us to construct infection trees for each simulation. We are also able to obtain metadata about individual simulations from the simulations dataset. This includes information about a variety of simulation parameters such as the total population size, the number of initial infections, the duration of the simulation, and the pathogen being spread (e.g., SARS-CoV-2, Influenza Virus, and others). It also captures environmental and intervention settings, such as whether location tracking and beacons were used, and the structure of protective measures like mask usage and Personal Protective Equipment (PPE). The pathogens dataset provides detailed characteristics of the infectious agents modeled in the simulations, linking directly to the simulations dataset via the pathogen ID. Each pathogen entry defines key epidemiological parameters, including transmission probabilities, incubation and symptomatic periods, severity progression, and the effectiveness of two specified levels of protective measures: wearing a mask and wearing PPE.

2.2. How was the data prepared?

We start with a total of 72 Operation Outbreak simulations, each with a unique ID number and a name (for example, simulation 165 is named 'WKU Outbreak' for Wenzhou-Kean University where a large simulation was run [8]). Many of these simulations, and some of the users participating in these simulations, are not eligible to be included in our analyses. We first discard users that were not active. We define an active user to be a user that is part of at least one contact event over the duration of a simulation. Next, to incorporate information on protection status (no protection, mask, PPE), we turn to 'modifier' events within the histories dataset, as indicated by the 'type' column. We track this status by adding an additional column that keeps track of the user's current protection status for each row or event in the dataset. Given our focus on 'infection' events, as specified by the 'type' column in the histories dataset, our next step in preparing the data involved isolating these events from the complete histories dataset. With some additional cleaning of data found in the Appendix Section A, this leaves us with a dataset of infection events for 44 of the 72 original simulations.

2.3. Network Analysis Tools

The primary tool that we used for network analysis and visualization is the NetworkX Python package. This package is commonly used for network analysis and to implement network algorithms. NetworkX supports data structures including many types of graphs and directed graphs which are central to network analysis. The use of Python for NetworkX makes our analysis more accessible for students and non-experts [9]. Using this package, we represent infected individuals as nodes and infection events as edges in a directed graph. Additionally, packages such as pandas, NumPy, and SciPy are used for data manipulation and numerical computations, and Matplotlib and Seaborn are used for data visualization.

2.4. Infection Trees

A key component of our analysis is the construction and examination of infection trees for each simulation. These trees represent the chain of transmission, where each node corresponds to an infected individual and each directed edge represents a transmission event. Nodes are colored based on protection status at the time of infection:

- **Gray:** Initial case
- **Red:** No protection
- **Blue:** Mask
- **Green:** PPE

Each simulation produces one or more complete infection trees that capture all observed transmissions. Infections may occur in isolated clusters that are not connected, since multiple "patient zero" infections are included in any given simulation. This reflects a realistic scenario, where a larger outbreak in the wild could be introduced to a local community by multiple individuals.

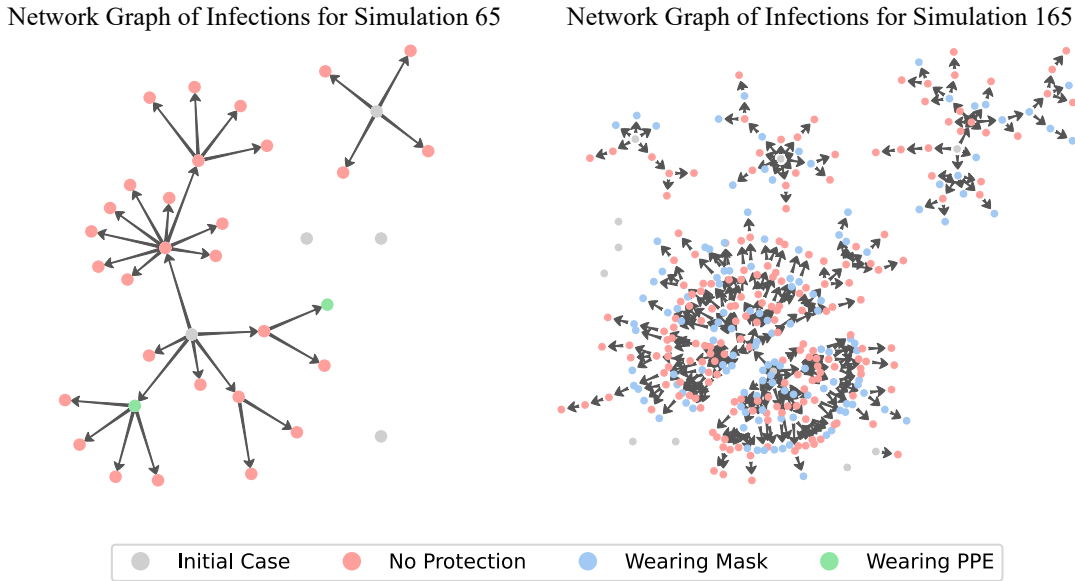


Figure 1: Infection trees for Simulations 65 (Haughton) and 165 (WKU Outbreak). Nodes represent infected individuals, with directed edges indicating transmission events. Colors denote protection status: gray (initial case), red (no protection), blue (mask), and green (PPE).

3. Evaluating the Realism of Operation Outbreak Simulations

Our main research question is if the infection trees generated through Operation Outbreak simulations resemble real-world transmission dynamics. If the transmission dynamic of these trees aligns with those observed in actual outbreaks, this would validate the utility of Operation Outbreak as a model for studying disease spread. In contrast, significant discrepancies might indicate limitations in the digital transmission model.

To explore this question, we follow a similar approach to Taube et al. [5] in their OutbreakTrees study, and benchmark the structural properties of infection trees from Operation Outbreak simulations against these real-world transmission networks. Detailed parameter estimates and summaries for each simulation studied in this section can be found in Appendix Section B.

3.1. Parameter Estimation

We will be estimating common network and infection parameters. The first parameter of interest is the basic reproductive number R_0 , or the average number of secondary transmissions per case. We will also estimate the dispersion parameter k of a negative binomial distribution fit to the offspring distribution of trees.

Hébert-Dufresne et al. [4] detail the importance of accounting for heterogeneity when modeling disease spread, presenting a negative binomial network model through the assumption that the distribution of secondary infections follows a negative binomial distribution. The number of secondary infections is not fixed at R_0 , as it can vary significantly between individuals. Some individuals transmit the disease to many others, while others transmit it to few or none. The negative binomial model serves as a useful parametric summary of this variation, capturing this overdispersion in transmission using two parameters, R_0 and k . This concept of heterogeneity is further supported by research done by Hébert-Dufresne et al. [10] into forecasting disease spread for the 2014-2016 Ebola Virus Disease epidemic in Sierra Leone. They found that most cases in an outbreak produce a smaller than average number of secondary infections, with a very small proportion (around 1 in 1000, in this case) leading to orders of magnitude more infections, highlighting the importance of understanding the heterogeneity of disease spread.

We estimate R_0 and k using maximum likelihood estimation of a negative binomial distribution. This distribu-

tion is fit to the offspring distribution of trees, where R_0 is the average number of secondary transmissions per case and k is the dispersion parameter. The use of maximum likelihood estimation for estimating k is discussed by Lloyd-Smith [11], where it is explained how the use of the negative binomial distribution to model secondary infections is motivated by its ability to account for overdispersion, capturing the variability in transmission where some individuals cause many secondary cases while others cause few or none. According to the negative binomial model, the number of secondary cases X caused by an infection is a random variable with pmf

$$\Pr(X = x) = \frac{\Gamma(x + k)}{x! \Gamma(k)} \left(\frac{R_0}{R_0 + k} \right)^x \left(1 + \frac{R_0}{k} \right)^{-k}, \quad (1)$$

parameterized by reproductive number R_0 , which is the mean of the distribution, and dispersion parameter k . Due to the flexible framework for modeling overdispersed count data, the negative binomial model is well suited to estimate R_0 and k from secondary infection data. The maximum likelihood estimation (MLE) of parameters for the negative binomial distribution is explained by Piegorsch [12], who shows that the MLE can be obtained by maximizing

$$\ell(k, R_0) = \frac{1}{n} \sum_{i=1}^n \sum_{v=0}^{y_i-1} \left(\log \left(1 + \frac{v}{k} \right) \right) + \bar{y} \log R_0 - (\bar{y} + k) \log \left(1 + \frac{R_0}{k} \right), \quad (2)$$

which is a log-likelihood function defined by modeling each infection case as independent and identically distributed (i.i.d.) events x_1, x_2, \dots, x_n , where n is the total number of cases, and y_1, \dots, y_n are secondary infection counts for each case. Using MLE with the log-likelihood function from Equation 2, we estimate R_0 and k for each simulation to be the values that minimize the negative log-likelihood function using the L-BFGS-B (Limited-memory Broyden-Fletcher-Goldfarb-Shanno with Box constraints) algorithm [13]. Estimation of parameters was performed for simulations with a minimum of 20 secondary infections, since simulations with only a small number of secondary infections had difficulty converging through MLE ($k \rightarrow \infty$). There were eight simulations that fit this criterion.

The negative binomial model appears to provide a good fit for the distribution of secondary infections, as shown in the goodness-of-fit analysis in Appendix Section C. The figures compare the distributions for simulations using estimated parameters \hat{R}_0 and \hat{k} to the observed data, showing a close fit that supports findings such as those discussed by Lloyd-Smith [11], who noted that maximum likelihood estimates of k are reliable for highly overdispersed datasets like those seen in disease transmission.

To further illustrate the likelihood landscape of our estimates, we created a heatmap of the log-likelihood estimates as a function of R_0 and k for the WKU Outbreak simulation, the largest simulation, as seen in Figure 2. The heatmap shows a single, moderately sharp peak around the MLE, indicating that the parameter estimates for R_0 and k are fairly well-constrained. The concentration of the bright region near the MLE suggests that these estimates are reasonably robust. While there is some uncertainty, the likelihood surface provides good support for the chosen parameter values. Because the heatmap reflects the parameter estimates for the full outbreak tree, the true R_0 must be less than one. However, this still effectively illustrates the landscape of our parameter estimates. Later, in Section 3.3, we present another likelihood landscape, this time visualizing parameter estimates based on only the first half of the simulation.

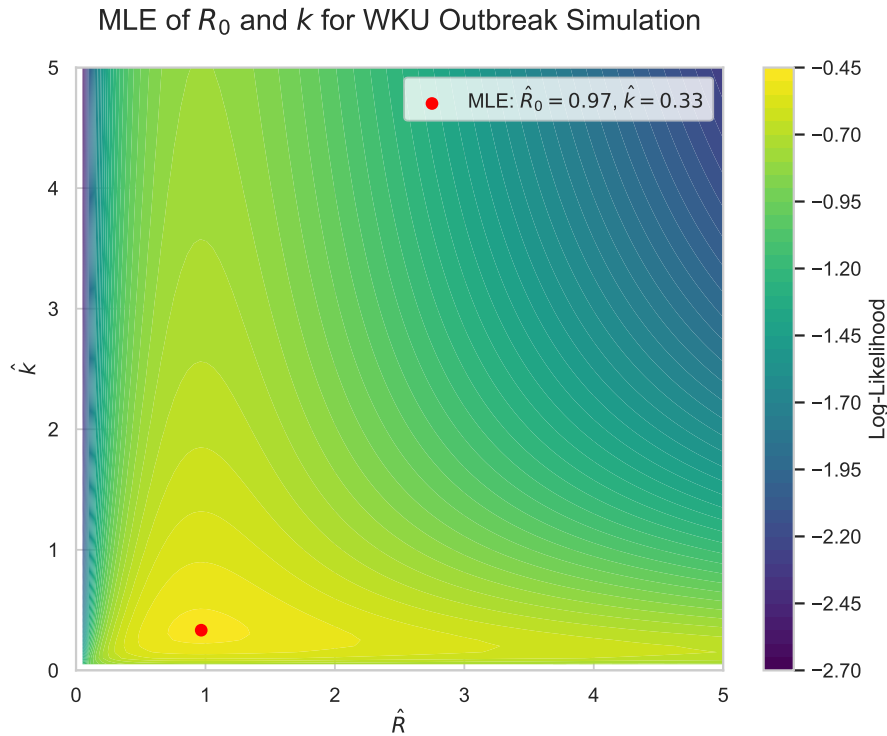


Figure 2: Heatmap of the log-likelihood surface for the WKU Outbreak simulation. Darker regions of the surface indicate lower likelihood values and the brightest region indicates the maximum likelihood estimate. The red dot marks the estimated parameters, \hat{R}_0 and \hat{k} .

Comparing our results for parameter estimations to OutbreakTrees, we observe some differences in the range of estimates, particularly in our estimates of \hat{R}_0 . Tables 1 and 2 compare our parameter estimates to estimates found in data tables from OutbreakTrees [14] from 39 infection trees that contain at least 20 cases and at least two generations of spread.

Percentile	OutbreakTrees Including Terminal Nodes	Operation Outbreak Including Terminal Nodes
Minimum	0.95	0.05
25th Percentile	0.96	0.31
Median	0.97	0.64
75th Percentile	0.98	0.89
Maximum	1.00	0.99

Table 1: Comparison for quantiles of R_0 estimations between OperationOutbreak and OutbreakTrees.

Percentile	OutbreakTrees Including Terminal Nodes	Operation Outbreak Including Terminal Nodes
Minimum	0.020	0.074
25th Percentile	0.085	0.126
Median	0.160	0.183
75th Percentile	0.580	0.320
Maximum	100.000	0.333

Table 2: Comparison for quantiles of k estimations between OperationOutbreak and OutbreakTrees.

However, when comparing these results, it is important to consider the fundamental differences in these datasets. An Operating Outbreak simulation may include many smaller transmission chains and initial infections that do not go on to spread or cause larger outbreaks. OutbreakTrees, however, performs parameter estimation only on trees with a single initial infection, at least 20 infections per tree, and at least 2 generations of spread.

As expected, the inclusion of these smaller chains and isolated initial infections dragged down estimates of R_0 compared to the real-world trees. Estimates of dispersion parameter k also spanned a smaller range for Operation Outbreak than with OutbreakTrees. Despite these differences, which we do not believe indicate a limitation of Operation Outbreak, we were able to successfully fit the offspring distribution to a negative binomial model, which appears to be a good fit for our data. This is consistent with previous studies on infectious disease transmission, and it opens the door for us to perform additional analysis such as parameter-time dependence to see how these parameters change as a disease spreads.

3.2. Superspreader Analysis

Lloyd-Smith et al. [7] define superspreader events as being in the 99th percentile of the $Poisson(R_0)$ distribution of expected secondary infections. This definition means that superspreader events are very unlikely to occur under a model of spread with no overdispersion, where most individuals cause a similar number of secondary infections close to R_0 . This makes superspreaders stand out as being individuals who transmit the disease to many more people than expected under a no-overdispersion model.

For all superspreader analysis, we will use the definition of superspreaders as outlined by Lloyd-Smith et al. We did some experimenting with this definition, exploring other percentile cut-offs for the "superspreader" label. For simulation WKU Outbreak, Figure 3 shows the proportion of cases that are considered "superspreaders" at various percentile cutoffs. We decided to stick with the 99th percentile, removing any simulations where the definition of superspreader is an infection event leading to less than 3 secondary infections.

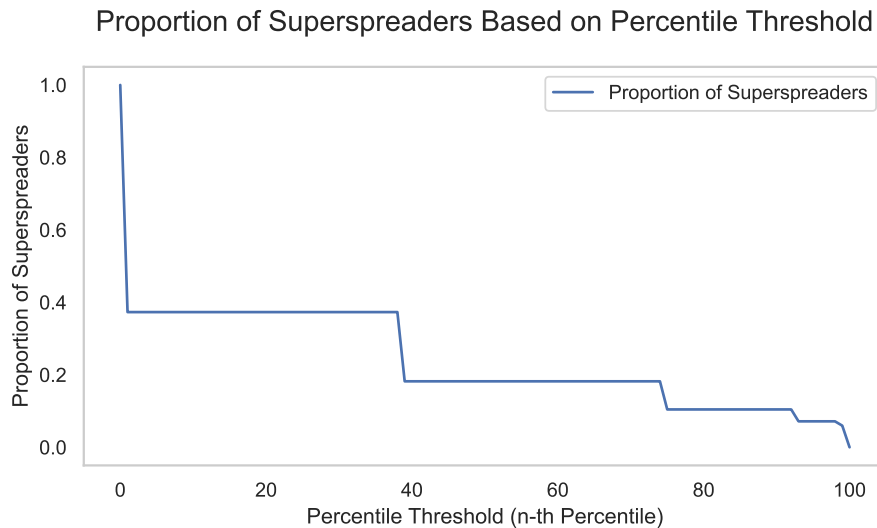


Figure 3: Proportion of cases classified as superspreaders at various percentile cutoffs for simulation WKU Outbreak. The proportion of cases classified as superspreaders at the 99th percentile cutoff is 0.0597.

Analysis of OutbreakTrees [5] found that superspreaders infected other superspreaders at a higher-than-expected rate in two-thirds of infection trees examined. The expected number of superspreader-superspreader dyads in an infection tree is

$$E_{SS} = \frac{s(s-1)}{(S-t)}. \quad (3)$$

This equation, from the superspreader analysis conducted in OutbreakTrees, is based on combinatorics under a random transmission model, where s is the number of superspreaders, t is the number of terminal nodes, and S is the size of the tree. This calculates the number of possible pairs of superspreaders, and uses the tree size to determine how many of these pairs we would expect to see by chance (adjusting for the number of terminal infectors in the tree). To ensure consistency with the OutbreakTrees analysis, we restricted our analysis to simulations that contained at least one superspreader and at least 20 secondary infections. The

calculated ratios of observed to expected superspreader-superspreader dyads for these simulations is displayed in Figure 4.

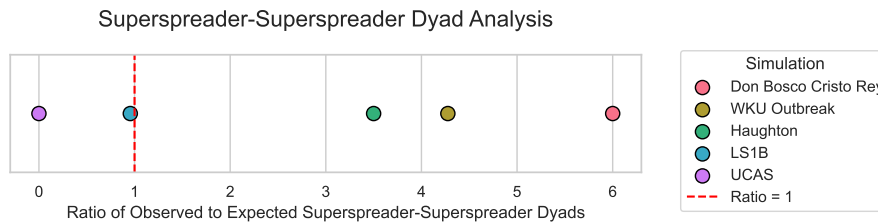


Figure 4: Ratio of observed to expected superspreader-superspreader dyads for simulations containing at least one superspreader and at least 20 secondary infections. Ratios greater than one (marked by the dashed red line) indicate that superspreaders infected other superspreaders at a higher-than-expected rate in that simulation.

Our findings align closely with those of OutbreakTrees. We observed that 3 out of 5 (60%) of the eligible simulations had an observed-to-expected superspreader-superspreader dyad ratio greater than one. The highest ratio was found in the Don Bosco Cristo Rey simulation, where the observed number of dyads was six times the expected value. This simulation contained 9 superspreaders and 8 superspreader-superspreader dyads, compared to an expected 1.33 dyads.

3.3. First vs. Second Half of Tree

This section of our research explores how parameter estimations of R , k , and the number of superspreaders changes throughout a simulation. We will do this by splitting each simulation into first and second halves by time of infection, picking a split time such that there is an equal number of cases in either half of the simulation. If there is an uneven number of cases in a simulation, the extra case will be assigned to the first half of the simulation. This adapts the method of splitting trees into first and second halves used by Taube et al. [5] in the OutbreakTrees study, where trees were split by generation. Because we have access to infection timestamps within the Operation Outbreak simulations, however, we decided that this would improve the accuracy of our analyses into how a disease changes over time.

In earlier analyses, we used R_0 to refer to the reproduction number over the entire outbreak. When splitting the outbreak into halves, we use R to refer to reproduction numbers estimated separately for each half. Figure 5 shows the likelihood landscape for the first half of the WKU Outbreak simulation, highlighting the parameter estimates based on the early-stage transmission dynamics.

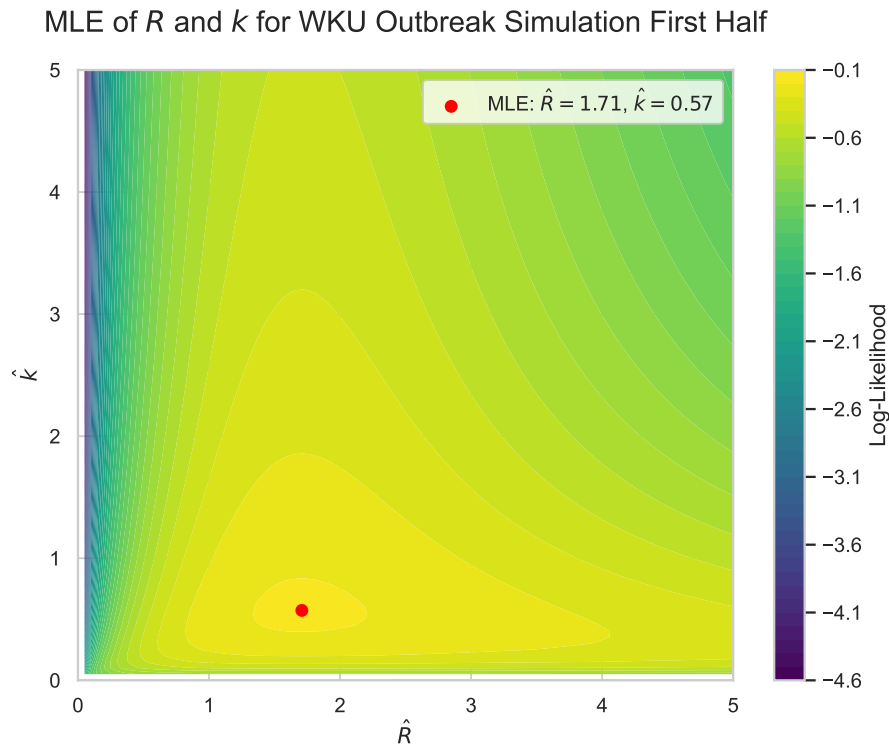


Figure 5: Heatmap of the log-likelihood surface for the first half of the WKU Outbreak simulation. Darker regions of the surface indicate lower likelihood values and the brightest region indicates the maximum likelihood estimate. The red dot marks the estimated parameters, \hat{R} and \hat{k} .

To investigate how transmission dynamics evolve throughout an outbreak, we compare the estimates of R and k between the first and second halves of each simulation. By plotting the estimates for each half in a shared 2D space, we visualize shifts in transmission patterns over time. Figure 6 illustrates these changes in the five simulations that met our inclusion criteria, had a minimum of 20 secondary infections, and had converging estimations of R and k .

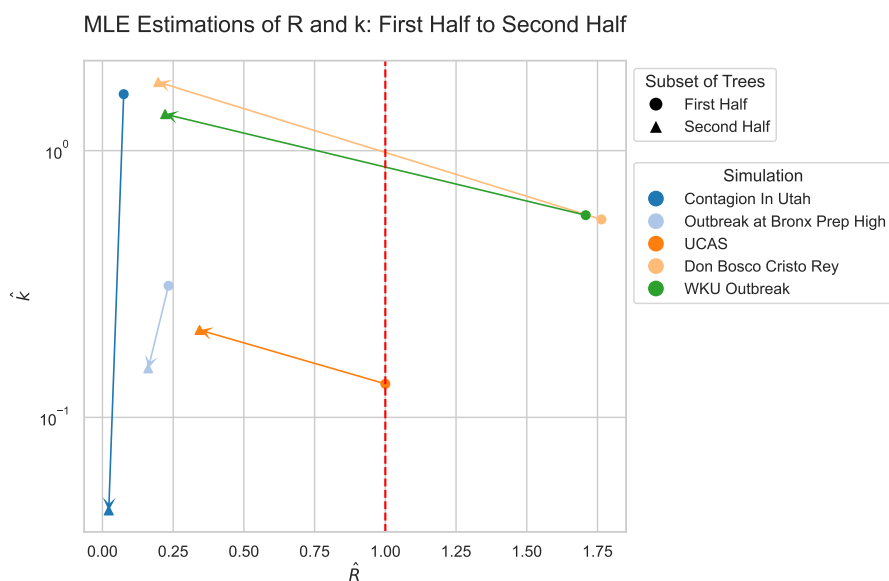


Figure 6: Visualization of changes in estimated R and k between the first and second halves of five simulations. Each point represents the maximum likelihood estimations for disease parameters for one half of a simulation, with arrows indicating the direction of change from the first half to second half of the simulation. The dotted red line represents $R = 1$.

Across all simulations, the maximum likelihood estimate (MLE) of R for the second half of the simulation was consistently less than 1, reflecting that the outbreak has reached the extinction phase. Every simulation also showed a decrease in R from first half to second half. In contrast, the dispersion parameter k showed no clear directional trend. While three simulations had an increase in k , suggesting reduced superspreading, two simulations showed a decrease, indicating greater transmission heterogeneity in the later stages of the outbreak.

Next, we look at how the number of superspreaders changes from the first to the second half of the simulations. We observe a substantial decrease in the number of superspreaders in the four simulations which meet the criteria of having superspreader cutoff ≥ 3 secondary infections and at least 20 infections as shown in Figure 7.

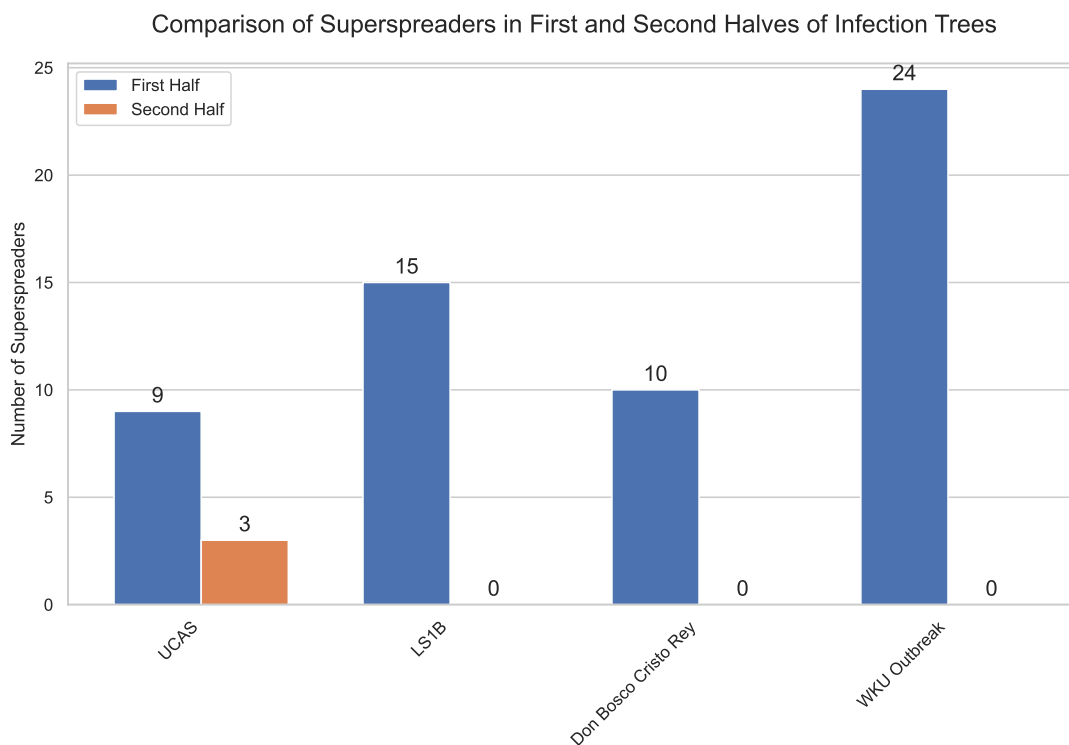


Figure 7: Change in the number of superspreaders between the first and second halves of infection trees. Each pair of bars represents one simulation, with the blue bar (left) representing the number of superspreaders in the first half of the simulation and the orange bar (right) representing the number of superspreaders in the second half of the simulation.

Our results from analyzing the split of these infection trees is mostly consistent with the results demonstrated by Taube et al. [5]. In the OutbreakTrees study, a decrease in R , increase in k , and decrease in the proportion of cases causing superspreading events were observed. Similarly, our analysis showed a decrease in R and a decrease in the proportion of cases causing superspreading events. However, we were unable to draw a clear parameter-time dependence for the dispersion parameter k . This may be due to the limited number of large simulations available for us to analyze. Although our results on superspreaders are consistent with the OutbreakTrees study, it is possible that superspreaders are more likely to appear in the first half of an outbreak, as individuals who generate many secondary cases could contribute disproportionately to early outbreak growth. Although our results on superspreaders are consistent with the OutbreakTrees study, it is possible that being a superspreader is statistically correlated with appearing earlier in the outbreak. Individuals who generate many secondary cases contribute more heavily to early outbreak growth, increasing the number of downstream cases, which may make it more likely for them to fall within the first half of the outbreak. To determine whether our results can be explained by this correlation, a comparison to a null model would be necessary.

4. Overcoming Limitations Through Operation Outbreak

We believe that Operation Outbreak agrees with real-world infection trees like those seen in the analysis of OutbreakTrees. Therefore, under the assumption that Operation Outbreak is a valid way of generating simulated infection data, this unlocks significant potential in the field of disease epidemiology. Since Operation Outbreak generates realistic infection trees, we can leverage the abundance of data from its simulations to conduct analyses that would be impossible in real outbreaks, where monitoring is far more limited.

Consider OutbreakTrees, where the following limitations were listed:

- Not a random or representative sample of directly transmitted infectious disease outbreaks.
- A transmission tree is typically an incomplete part of a larger chain of transmission events.
- Information about how control measures and behavior changes alter disease parameters in the middle of an outbreak is not available.

Through Operation Outbreak simulations, we can start to address the limitations that are unavoidable when studying real-world diseases. For example, we know that the transmission trees present in these simulations are not part of a larger chain of transmission events, as each simulation is self-contained and we have information on every single case in the study. Furthermore, we know about control measures that are in place through a user’s protection status, and we can even track behavior changes for participants in the middle of an outbreak through contact data.

There is so much potential when it comes to Operation Outbreak simulations, and we will briefly give a couple of examples of analyses that would be unfeasible to conduct on a large scale with real-world infection trees.

4.1. Mixing Pattern Analysis

Protection status for participants is tracked throughout an Operation Outbreak simulation. For mixing pattern analysis, we start by creating network graphs of individual simulations with nodes colored by protection status at the time of infection (initial case, no protection, wearing mask, wearing PPE) to visualize protection status within an infection tree. Next, across all simulations that fit our inclusion criteria outlined in Appendix Section A, we count the transition motifs to see how many edges (infection events) connect individuals of each protection level.

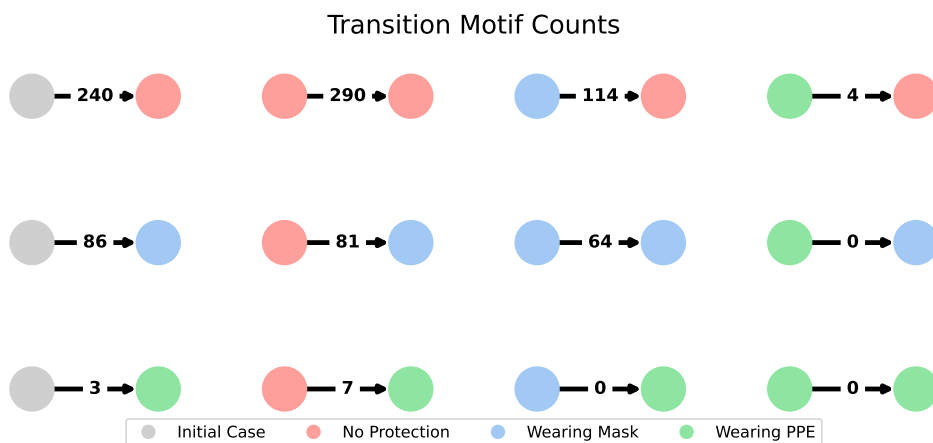


Figure 8: Transition motif counts representing the number of infection events between individuals of each protection level across all simulations that met the inclusion criteria at the time of infection.

From these data, we may then want to evaluate the tendency of individuals in an infection network to be grouped closely with others of the same protection status. We can evaluate this by calculating an assortativity value for each simulation. Assortativity values closer to 1 indicate that nodes with the same protection status are likely to be connected, values closer to -1 indicate that nodes with different protection statuses are likely

to be connected, and values around 0 indicate that nodes of the same protection status are not more or less likely to infect others of the same or different protection status. The equation used for directed assortativity, as described by Leicht and Newman [15], is

$$Q = \frac{1}{m} \sum_{i,j} \left(A_{ij} - \gamma \frac{k_i^{out} k_j^{in}}{m} \right) \delta(c_i, c_j). \tag{4}$$

Assortativity coefficient Q measures the tendency of nodes with the same protection status to be connected in the infection network, m is the total number of edges, A_{ij} is the adjacency matrix, k_i^{out} is the out-degree of node i , k_j^{in} is the in-degree of node j , γ is the resolution parameter (set to 1 for our use), and $\delta(c_i, c_j)$ is 1 if i and j have the same protection statuses c_i and c_j and 0 otherwise.

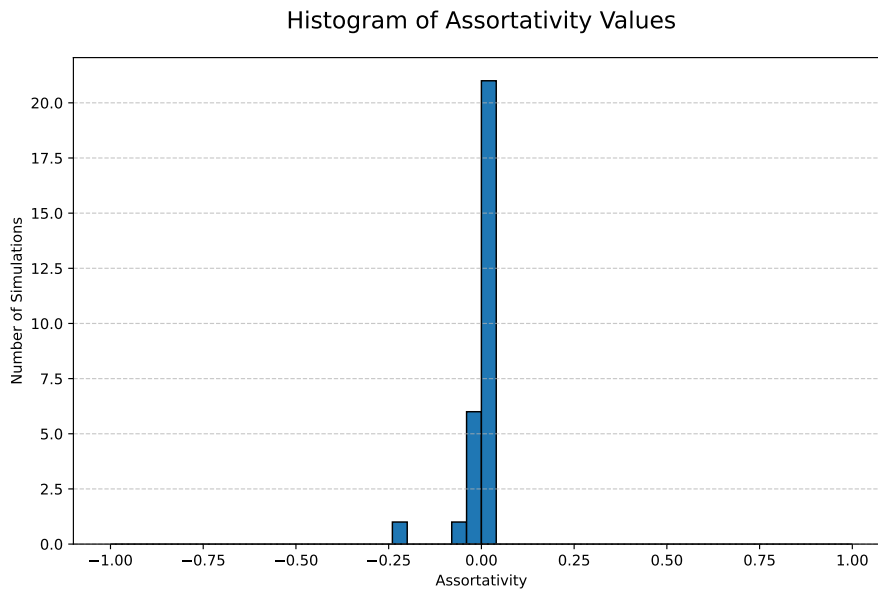


Figure 9: Histogram of assortativity values across all Operation Outbreak simulations that met the inclusion criteria.

As seen in Figure 9, assortativity values are centered closely around zero. This indicates that, in general, individuals are not more or less likely to infect others with a similar protection status.

This mixing pattern analysis would be impossible to perform with real-world infection trees. Control measures such as protection status are rarely known (especially for both the infector and the infectee) at the time of infection. Operation Outbreak simulations allow for these control measures to be tracked at any point during the simulation, providing a unique opportunity to study how protection status influences transmission dynamics that are unattainable without simulated data like that from Operation Outbreak.

4.2. Serial Intervals

Every event, whether it be a contact, infection, change in protection status, etc., that takes place during an Operation Outbreak simulation is tracked with a timestamp. This means that we know down to the second when someone is infected, who they were infected by, and how long that person has been infected for. Using this information, we can calculate serial intervals. A serial interval is the time between a person being infected and them infecting others. For the WKU Outbreak simulation, we calculate an individual’s serial interval to be the mean amount of time between being infected and passing the infection on to their secondary infections. For example, an individual who infects one person at $t = 50$ seconds and another person at $t = 100$ seconds will have a serial interval of $\frac{100+50}{2} = 75$ seconds between infection and transmission. We calculate this for each individual that causes at least 1 secondary infection. Figure 10 shows the distribution of mean serial intervals for each individual that causes at least 1 secondary infection in simulation WKU Outbreak.

Histogram of Serial Interval per Infection for WKU Outbreak Simulation

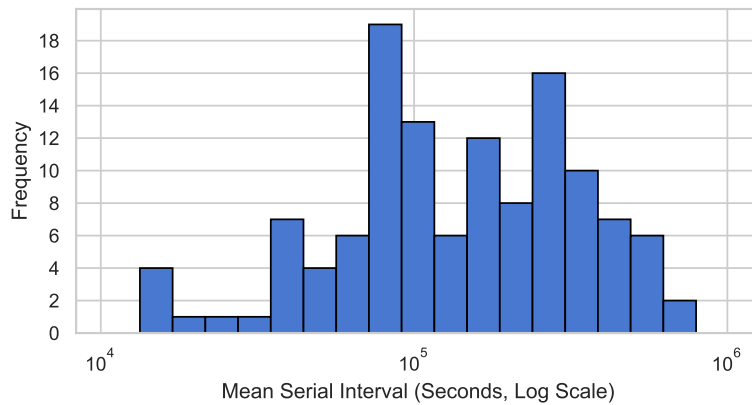


Figure 10: Histogram of serial interval values for simulation WKU Outbreak. A participant’s serial interval is calculated as the mean amount of time between them becoming infected and infecting others.

To analyze how these serial intervals change over time, we visualize the serial intervals for this simulation’s first and second halves in Figure 11.

Serial Intervals Across WKU Outbreak Simulation: First Half vs. Second Half

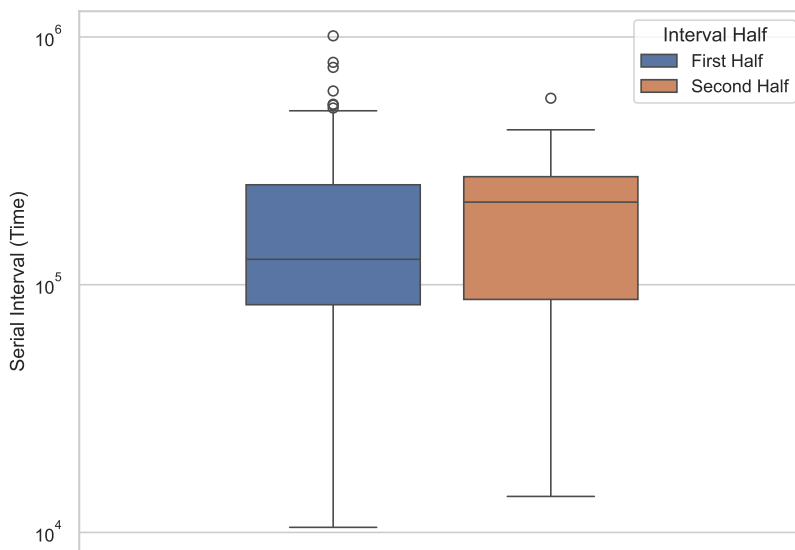


Figure 11: Side-by-side boxplots of serial interval values (in seconds) for simulation WKU Outbreak. The boxplot on the left shows the serial interval distribution for the first half of the simulation, the boxplot on the right shows the serial interval distribution for the second half of the simulation.

We calculate the overall median serial interval for simulation WKU Outbreak to be about 40.49 hours. An interesting relationship that we observe for this simulation is that the median serial interval increases from the first to the second half of the simulation, from 35.17 hours to 59.86 hours. To see if this is unique to this simulation, or if this is a trend that we observe in other simulations as well, we check this for a larger subset of 7 simulations in Figure 12 that meet the inclusion criteria, have at least 20 secondary infections, and have at least 2 secondary infections in each half of the simulation.

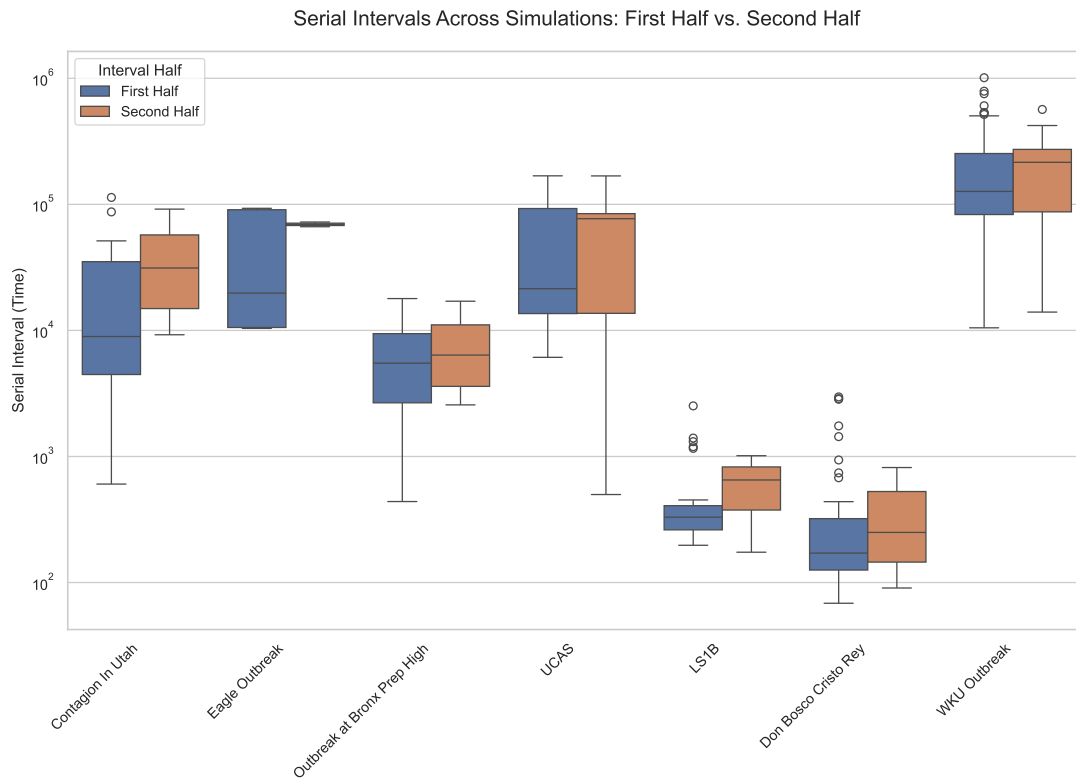


Figure 12: Side-by-side boxplots of serial interval values (in seconds) for simulation WKU Outbreak. The boxplot on the left shows the serial interval distribution for the first half of the simulation, the boxplot on the right shows the serial interval distribution for the second half of the simulation.

From this figure, we notice that the median serial interval increases in every simulation that meets these criteria. This is a very interesting trend that we may have missed if these data were subject to the limitations that real-world infection trees are. This pattern may suggest that as outbreaks progress, individuals take longer to transmit the pathogen, possibly due to increased awareness, changes in protection status, or some other factors. Further analysis could explore what may cause this trend.

4.3. Conditional Probability of Infection

Infection trees track who got infected during an outbreak, but not who stayed disease-free. In real-world outbreaks, that's often all you have — no perfect dataset of both infected and uninfected individuals to compare. Until now, in our analysis, we have focused on infection trees. But the simulations from Operation Outbreak captured more than just the spread of disease. They also recorded, for example, contacts and protection measures of those who didn't get infected, offering a broader view of how disease transmission played out.

Turning our attention to this broader view of disease transmission, we now have the opportunity to examine not just who was infected, but also who remained uninfected and why. To explore this further, we calculated the conditional probabilities of infection based on protection status — comparing those who wore PPE or masks to those who did not. This analysis provides insight into how protective measures influenced the likelihood of infection during the simulations.

In this analysis, we included all simulations that met the inclusion criteria outlined in Appendix Section A. This time, we expanded our focus to include all active individuals, not just those who were part of an infection event. We excluded primary infections, as our goal was to assess how the use of PPE or masks influenced peer-to-peer transmission. For simplicity, a user was classified as having worn PPE or a mask if they did so at any point during the simulation, or before becoming infected if applicable.

The results in Table 3 show a slight difference in infection rates based on protection status. Those who wore PPE

	P(Infected Protection Status)	P(Not Infected Protection Status)
Wore PPE/Mask	0.470	0.530
Did Not Wear PPE/Mask	0.494	0.506

Table 3: Conditional probabilities of infection and non-infection based on protection status for 1816 eligible participants. The table shows the likelihood of an individual becoming infected or remaining uninfected given whether they wore PPE or a mask during the simulation.

or masks had a 47.0% chance of becoming infected, while 53.0% remained uninfected. In contrast, individuals who didn't wear PPE or masks had a slightly higher infection rate of 49.4%, with 50.6% remaining uninfected. This suggests that wearing protection during these simulations may reduce the likelihood of infection, but the difference is relatively small compared to what we may expect. However, a potential confounding factor to consider is that individuals who wore PPE or masks may have had more interactions during simulations, which could have increased their chances of exposure to infection despite protective measures. This is something we are able to look into in the future with the available data provided by Operation Outbreak, as we have access to contact data other than transmissions.

Using simulated disease data like this allows us to explore the impact of protection measures on both infected and uninfected individuals by having access to protection status information about the entire population, something that we would not have access to with real-world diseases.

5. Conclusion

The results from our analysis indicate that Operation Outbreak simulations are a viable tool that can be used to gain insight into disease transmission dynamics, providing access to detailed infection data that would be extremely challenging, if not impossible, to obtain from real-world outbreaks. We benchmarked results for estimations of common disease parameters R_0 (the basic reproduction number) and k (the dispersion parameter of a negative binomial distribution), against the analysis of OutbreakTrees by Taube et al. [5]. Any differences, such as a larger range of values for R_0 , could potentially be explained by the nature of our data and factors such as its inclusion of multiple transmission chains and primary infections that go unspread. We then analyzed superspreaders and found that, similar to the results from OutbreakTrees, 60% of the eligible simulations had an observed-to-expected superspreader-superspreader dyad ratio greater than one. Additionally, when splitting simulations into first and second halves by timestamp, we observed a decrease in R from first half to second half in every simulation included in this part of the analysis. We also found that the proportion of cases causing superspreading events decreases, again consistent with real-world findings.

While further side-by-side comparisons will be needed to fully prove the viability of outbreak simulations as a tool for obtaining clean, reliable, and realistic infection data, this analysis offers promising evidence that Operation Outbreak simulations can capture key patterns of disease transmission and superspreading dynamics. This opens the door to disease analysis and research that would be nearly impossible without these simulations. In exploring this analysis and research that Operation Outbreak makes possible, we evaluated the tendency of individuals in an infection network to cluster with others of the same protection status using assortativity, but found no strong evidence for this pattern. We then analyzed serial intervals, finding that the time between becoming infected and infecting another participant increased from the first half to the second half of the infection in all simulations that we conducted this research on. Finally, we computed conditional probabilities for infection rates based on protection status, utilizing data from the entire population's contacts and infections. This revealed a slight decrease in the likelihood of an individual being infected given that they wore PPE or a mask during the simulation.

This research only scratches the surface of what is possible with Operation Outbreak simulations. Looking at the entire population, rather than just a single infection tree, opens up many new possibilities for future research [16]. We can continue exploring how different protection measures influence transmission dynamics, potentially setting up simulations that are geared towards improving our understanding of how changing protection methods during an outbreak may influence transmission patterns on a larger scale. Another possibility is to calculate infection curves, and we can utilize contact data from the population to evaluate the probability of infection based on the number of contacts an individual had [17]. The infection trees

themselves could also provide unique opportunities to study how a disease spreads in a controlled and realistic setting where we are able to collect clean and reliable data. This paves the way for future research that tests hypotheses about superspreaders, transmission dynamics, and other key factors that are difficult to study in real-world outbreaks. For example, future research could ask: Can we identify individuals who are at risk of becoming superspreaders before they are infected? How does targeted intervention on individuals with high contact counts alter the course of an outbreak? What is the effectiveness of interventions at different stages of an outbreak? Do people act more risky (have more contacts) when wearing a mask or wearing PPE? With Operation Outbreak simulations as a tool for studying disease transmission dynamics, we can better our understanding of how diseases spread and potentially inform responses to real-world outbreaks.

References

- [1] *Operation Outbreak*. en-US. URL: <https://operationoutbreak.org> (visited on 04/11/2024).
- [2] Leon Danon et al. “Networks and the Epidemiology of Infectious Disease”. en. In: *Interdisciplinary Perspectives on Infectious Diseases 2011* (Mar. 2011). Publisher: Hindawi, e284909. ISSN: 1687-708X. DOI: 10.1155/2011/284909. URL: <https://www.hindawi.com/journals/ipid/2011/284909/>.
- [3] Alun L. Lloyd and Steve Valeika. “Network models in epidemiology: an overview”. In: *Complex Population Dynamics*. Sept. 2007, pp. 189–214. DOI: 10.1142/9789812771582_0008. eprint: https://www.worldscientific.com/doi/pdf/10.1142/9789812771582_0008. URL: https://www.worldscientific.com/doi/abs/10.1142/9789812771582_0008.
- [4] Laurent Hébert-Dufresne et al. “Beyond R0: heterogeneity in secondary infections and probabilistic epidemic forecasting”. eng. In: *Journal of the Royal Society, Interface* 17.172 (Nov. 2020), p. 20200393. ISSN: 1742-5662. DOI: 10.1098/rsif.2020.0393.
- [5] Juliana C. Taube, Paige B. Miller, and John M. Drake. “An open-access database of infectious disease transmission trees to explore superspreader epidemiology”. en. In: *PLOS Biology* 20.6 (June 2022). Publisher: Public Library of Science, e3001685. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3001685. URL: <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001685>.
- [6] Ivan Specht et al. “Analyzing the impact of a real-life outbreak simulator on pandemic mitigation: An epidemiological modeling study”. In: *Patterns* 3.8 (Aug. 2022), p. 100572. ISSN: 2666-3899. DOI: 10.1016/j.patter.2022.100572. URL: <https://www.sciencedirect.com/science/article/pii/S2666389922001830>.
- [7] J. O. Lloyd-Smith et al. “Superspreading and the effect of individual variation on disease emergence”. en. In: *Nature* 438.7066 (Nov. 2005). Number: 7066 Publisher: Nature Publishing Group, pp. 355–359. ISSN: 1476-4687. DOI: 10.1038/nature04153. URL: <https://www.nature.com/articles/nature04153>.
- [8] Salihu S. Musa et al. “The effect of Behavioral Factors and Intervention Strategies on Pathogen Transmission: Insights from a Two-Week Epidemic Game at Wenzhou-Kean University in China”. In: *medRxiv* (Dec. 2024). DOI: 10.1101/2024.12.14.24318955. eprint: <https://www.medrxiv.org/content/early/2024/12/20/2024.12.14.24318955.full.pdf>. URL: <https://www.medrxiv.org/content/early/2024/12/20/2024.12.14.24318955>.
- [9] Aric Hagberg, Pieter J. Swart, and Daniel A. Schult. “Exploring network structure, dynamics, and function using NetworkX”. In: Los Alamos National Laboratory (LANL), Los Alamos, NM (United States). Jan. 2008. URL: <https://www.osti.gov/biblio/960616>.
- [10] Laurent Hébert-Dufresne et al. *The network epidemiology of an Ebola epidemic*. Nov. 2021. arXiv: 2111.08686 [q-bio.PE]. URL: <https://arxiv.org/abs/2111.08686>.
- [11] James O. Lloyd-Smith. “Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases”. en. In: *PLOS ONE* 2.2 (Feb. 2007). Publisher: Public Library of Science, e180. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0000180. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000180>.
- [12] Walter W. Piegorsch. “Maximum Likelihood Estimation for the Negative Binomial Dispersion Parameter”. In: *Biometrics* 46.3 (1990). Publisher: [Wiley, International Biometric Society], pp. 863–867. ISSN: 0006-341X. DOI: 10.2307/2532104. URL: <https://www.jstor.org/stable/2532104>.
- [13] Richard H. Byrd et al. “A Limited Memory Algorithm for Bound Constrained Optimization”. In: *SIAM Journal on Scientific Computing* 16.5 (1995), pp. 1190–1208. DOI: 10.1137/0916069. eprint: <https://doi.org/10.1137/0916069>. URL: <https://doi.org/10.1137/0916069>.
- [14] Juliana C. Taube, Paige B. Miller, and John M. Drake. *An open-access database of infectious disease transmission trees to explore superspreader epidemiology*. Dryad, June 2022. DOI: 10.5061/dryad.nk98sf7w7. URL: <https://datadryad.org/dataset/doi:10.5061/dryad.nk98sf7w7>.
- [15] E. A. Leicht and M. E. J. Newman. “Community Structure in Directed Networks”. In: *Phys. Rev. Lett.* 100 (11 Mar. 2008), p. 118703. DOI: 10.1103/PhysRevLett.100.118703. URL: <https://link.aps.org/doi/10.1103/PhysRevLett.100.118703>.

- [16] Martijn H. H. Schoot Uiterkamp et al. *Value of risk-contact data from digital contact monitoring apps in infectious disease modeling*. 2025. arXiv: 2503.21228 [q-bio.PE]. URL: <https://arxiv.org/abs/2503.21228>.
- [17] Nicholas W. Landry et al. "Reconstructing networks from simple and complex contagions". In: *Physical Review E* 110.4 (Oct. 2024). ISSN: 2470-0053. DOI: 10.1103/physreve.110.1042301. URL: <http://dx.doi.org/10.1103/PhysRevE.110.L042301>.

A. Inclusion Criteria

A.1. Participant Inclusion Criteria

- Users must be *active* (involved in at least one contact event during the simulation, around 69.9% of users were active).
- Users cannot be infected by an individual who has not yet been infected.
- Users can only be infected by another user within the same simulation.

A.2. Simulation Inclusion Criteria

- Simulations must contain active users and at least one infection event (23 simulations excluded).
- Peer infections must correctly specify the infecting user (3 simulations excluded).
- Simulations with inconsistent infection events were excluded from analysis (2 simulations excluded).

A.3. Final Simulation Selection

After applying these criteria, 44 simulations were included in the analysis, while 28 were excluded.

B. Data Table

Sim ID	Sim Name	Pathogen	Num Infections	R_0	k	Median Serial Interval	Num Superspreaders	Superspreader Dyad Ratio	...
57	Contagion In Utah	SARS-CoV-2	450	0.0489	0.3265	13494.00	N/A	N/A	...
59	Eagle Outbreak	SARS-CoV-2	58	0.3448	0.0739	66455.00	N/A	N/A	...
61	Outbreak at Bronx Prep High	SARS-CoV-2	212	0.1981	0.2239	5896.00	N/A	N/A	...
65	Haughton	Influenza virus	36	0.8612	0.1005	19104.50	1	3.5000	...
88	UCAS	Measles virus	136	0.6765	0.1345	64223.00	0	0.0000	...
118	LS1B	Measles virus	206	0.5971	0.1420	336.00	2	0.9556	...
131	Don Bosco Cristo Rey	Staphylococcus aureus	143	0.9860	0.3181	175.80	8	6.0000	...
165	WKU Outbreak	Virus X	335	0.9672	0.3332	145762.00	13	4.2763	...

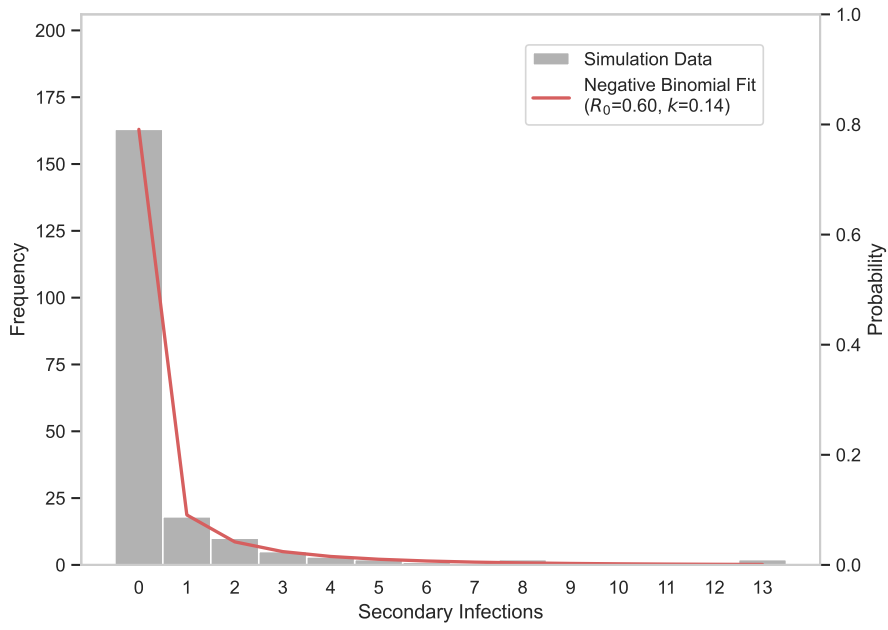
Sim ID	...	R First Half	R Second Half	k First Half	k Second Half	Superspreaders First Half	Superspreaders Second Half	Serial Interval First Half	Serial Interval Second Half
57	...	0.0752	0.0223	1.6287	0.0448	N/A	N/A	8945.50	31265.50
59	...	N/A	N/A	N/A	N/A	N/A	N/A	19778.50	69434.50
61	...	0.2336	0.1619	0.3115	0.1526	N/A	N/A	5499.00	6372.00
65	...	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
88	...	1.0000	0.3433	0.1336	0.2126	9	3	21426.33	76939.75
118	...	N/A	N/A	N/A	N/A	15	0	329.97	651.00
131	...	1.7639	0.1972	0.5523	1.8077	10	0	171.33	249.75
165	...	1.7083	0.2216	0.5729	1.3700	24	0	126629.75	215495.00

Table 4: Parameter estimates and summary information for Operation Outbreak simulations with at least 20 secondary infections.

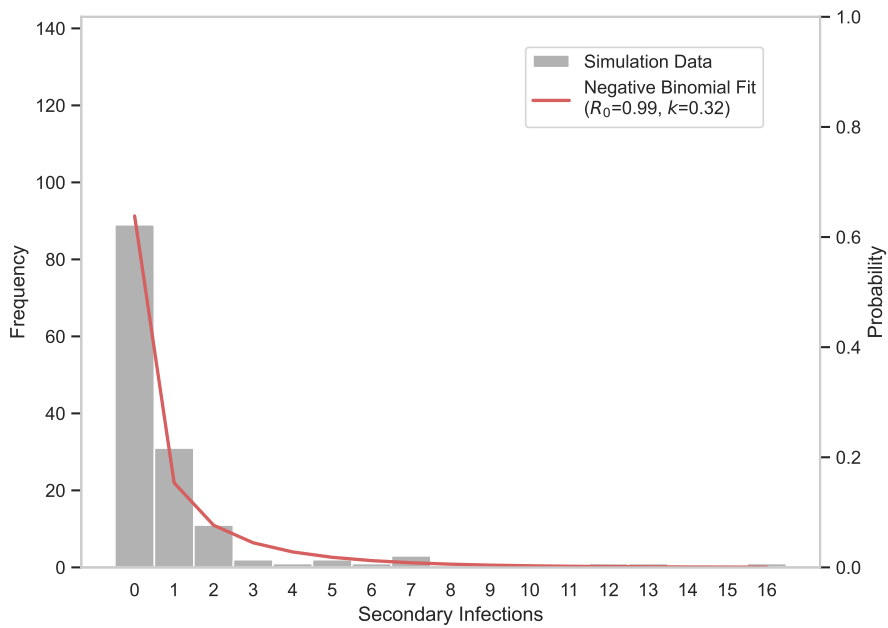
C. Goodness of Fit for Negative Binomial Distribution

This section presents a small collection of example simulations, visually comparing their distributions of secondary infections to negative binomial models fitted using the estimated disease parameters, \hat{R}_0 and \hat{k} . This comparison demonstrates how the negative binomial model appears to provide a good fit for the distribution of secondary infections.

LS1B Simulation: Fit of Negative Binomial Distribution



Don Bosco Cristo Rey Simulation: Fit of Negative Binomial Distribution



WKU Outbreak Simulation: Fit of Negative Binomial Distribution

