A Deep Dive into Recreating and Analyzing Superspreading Models

Michael Dempsey, David Katilius

1 Introduction

Overdispersion refers to when there is more variability in the data than expected based on a statistical model. Superspreading is a concept in which there is an overdispersion of secondary infections relative to a Poisson model. This overdispersion may occur due to a wide variety of reasons such as different contact rates, duration of infection, and infectiousness. When this overdispersion occurs, there are a small number of people who cause a large number of infections within a population.

According to a paper by Dr. Lloyd-Smith et al. [6], the authors describe a general way of defining superspreading events.

1) Estimate the effective reproductive number, R.

2) Produce a Poisson distribution with mean R, which shows the expected number of Z secondary infections

3) Define an SSE (Super Spreading Event) as any infected individual who infects more than $Z^{(n)}$ others, where $Z^{(n)}$ is the nth percentile of the Poisson(R) distribution.

Essentially, the authors state that when an individual infects others at a rate much higher than what is expected, they are to be classified as a superspreader.

With this definition of superspreading, a few questions spring up.

1) How can one model these events? A huge spike in the number of secondary infections can make it challenging to fit a model to the data.

2) How does a superspreading event affect a population? Does it provide weird repercussions like having a dominant variant of a disease come out early and lead the way for future variations?

To answer these questions, we hope to take a look at the three models proposed in the Lloyd-Smith et al. paper and recreate their results.

Then, we will take a closer look at the effects and unique scenarios that occur due to superspreading by doing a literature review.

In the Lloyd-Smith et al. paper [2], three models for superspreading are proposed. Consider a distribution of individual reproductive numbers, $Pr(v)$, with a mean value of R_0.

$$E[P\, r(v)] = R_0$$

In this case, v_i represents the typical number of secondary infections caused by person i. The actual number of infections caused by person i is found by drawing from a Poisson distribution that uses v_i as the mean.

$$Z_i \sim \text{Poisson}(v_i)$$

There were three cases that the paper considered.

1) (*Pr(v)* is δ) In the first case, $v = R_0$ and the offspring distribution is Z~Poisson(R_0).

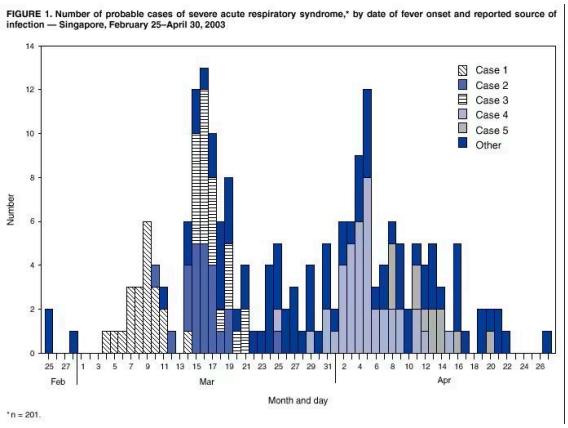
2) (*Pr(v)* is exponential) In the second case, v is exponentially distributed with mean $R_0 + 1$. The offspring distribution is represented by a geometric distribution, where $p = \frac{1}{R0+1}$, the mean is equal to R_0, and the number of secondary infections start at 0.

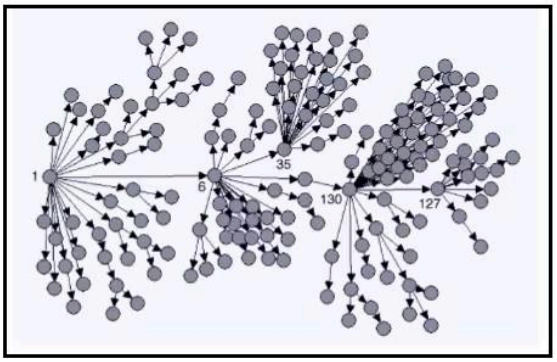
3) (*Pr(v)* is geometric) In the third case, v is Gamma distributed with mean R_0 and a dispersion parameter with value k. The offspring distribution is represented by a negative binomial distribution with $p = \frac{k}{k+R0}$, dispersion parameter equal to k, and mean value of R_0.

2 Methods

Our first step was to recreate some results from the Lloyd-Smith paper. We chose this paper because it is considered foundational for describing superspreading, meaning that all other papers which talk about superspreading events will have some elements similar to its contents.

The paper used numerous datasets to model superspreading. The first dataset they used involved SARS cases in Singapore back in 2003. The SARS dataset was provided on the CDC's website in the form of various diagrams (CDC 2003).





* Patient 1 represents Case 1; Patient 6, Case 2; Patient 35, Case 3; Patient 130, Case 4; and Patient 127, Case 5. Excludes 22 cases with either no or poorly defined direct contacts or who were cases translocated to Singapore and the seven contacts of one of these cases.
Reference: Bogatti SP. Netdraw 1.0 Network Visualization Software. Harvard, Massachusetts: Analytic Technologies, 2002.

Figure 1: The figures above show the SARS data collected in Singapore from February to April of 2003. Essentially, five different SARS cases were tracked, and the number of cases linked to the five were recorded. The first diagram shows the total number of cases each day. The cases were separated into categories based on which of the five cases they stemmed from. If a case could not be linked to any of the five, it was placed in the "Other" category. The second diagram shows a tree of the secondary cases linked to the five patients.

With only these two diagrams and no datasheets, we had to manually add the data into a spreadsheet. Once this was done, we were able to replicate the first diagram depicting the total number of cases each day. No additions or subtractions were made to the data.

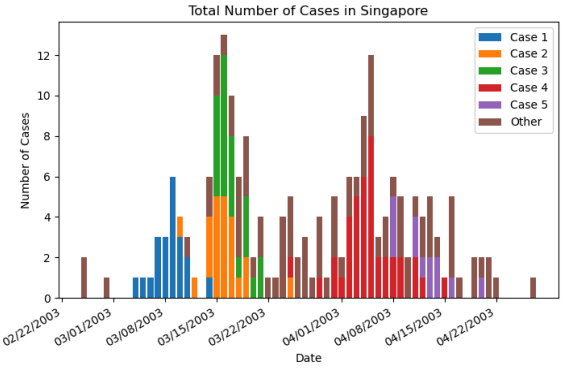


Figure 2: This is our replicated plot of the Singapore dataset, depicting the total number of cases each day.

For the case tree diagram, on the other hand, we did not add all of the data points. The paper put the first three generations together in their own dataset to represent the spread of infection before control measures were enacted. To account for this, we got rid of all the data points after the third generation and manually added the rest into a spreadsheet. We counted the number of secondary infections for each case and the frequency at which they occurred.

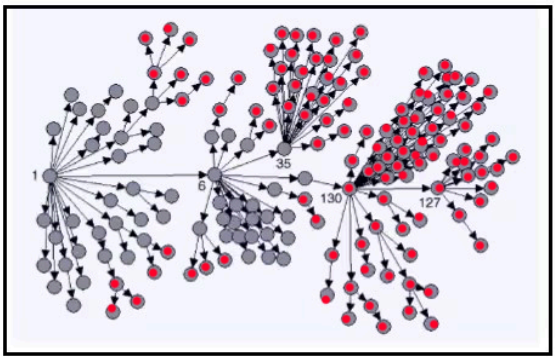


Figure 3: This is the case tree diagram from the Singapore dataset with red markings indicating cases after the third generation. We did not include any of these red cases in our analysis.

From the resulting spreadsheet, we were able to plot the frequency distribution of the secondary cases.

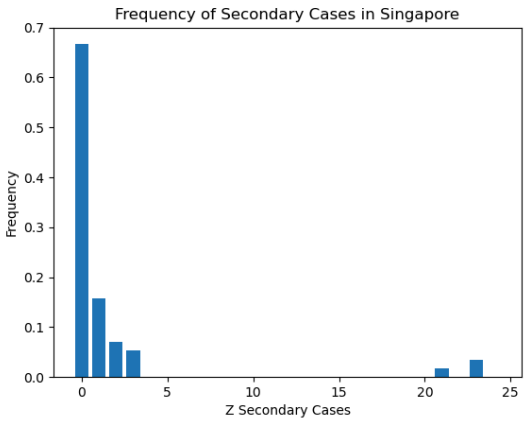


Figure 4: This is the frequency diagram of the number of secondary cases from the modified Singapore dataset (excluding all the cases after the third generation).

3 Results

After collecting the data used in the paper, we were able to start running simulations using the three models. First, we estimated the value of R_0. This was done by calculating the mean of the secondary cases, which ended up being approximately 1.632.

For the Poisson distribution, the values were calculated by passing in the frequency of the secondary cases and having μ be equal to R_0.

For the Geometric distribution, the values were calculated by finding p, which was equal to $\frac{1}{1+R0}$. The value of p ended up being approximately 0.38. Using the frequency of the secondary cases and p, the Geometric distribution was found.

For the Negative Binomial Distribution, the values were calculated by fitting the number of secondary cases to a Negative Binomial model, yielding two parameters. The first parameter was equal to the logarithm of the mean, so to calculate μ, $e^{log(\mu)}$ was found. The value of μ was correctly identified to be the same as the R_0 estimate we found previously with a value of 1.632. The second parameter, α, was equal to approximately 6.362. With these

values, the probability of success was calculated using the equation $\frac{1}{1+\mu\cdot\alpha}$. The probability of success had a value of approximately 0.0879. From here, the dispersion parameter k was found using the equation $\frac{\mu\cdot p}{1-p}$. The value of k ended up being approximately 0.157. The process of fitting the data and finding the parameters was aided by an article written by Anton Granik [5]. Using the p and k values and the secondary infection frequency, the Negative Binomial values were found.

We were able to plot the Poisson, Geometric, and Negative Binomial values on the frequency diagram.

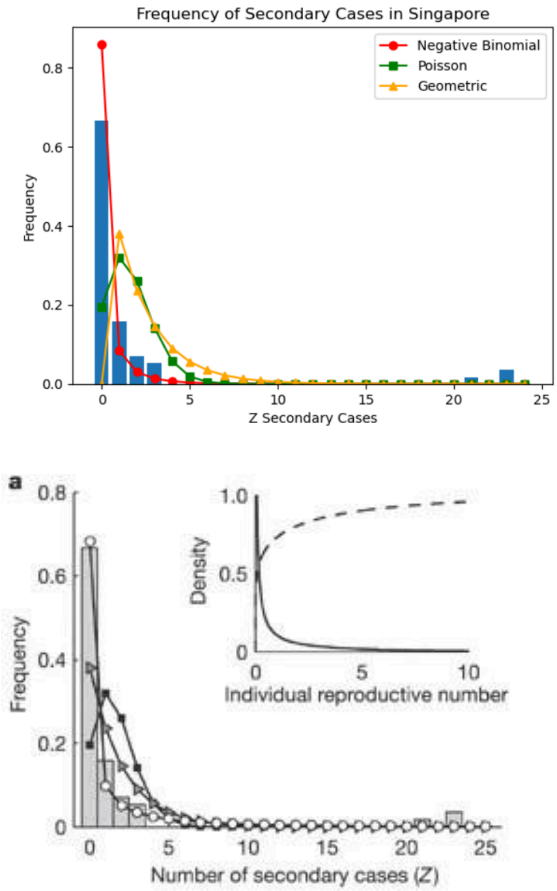


Figure 5: The top plot is our frequency diagram with the Poisson, Geometric, and Negative Binomial values plotted as well. The bottom diagram originates from the Lloyd-Smith paper. We were able to replicate the findings of the paper fairly accurately, with only some minor differences.

After plotting the data, we compared more of our results to the results of the paper. We were able to reproduce similar findings.

Parameter Estimation	SARS, Singapore (Loyd-Smith et al)	Our Calculations
N	57	57
R_0	1.63	1.632
K	0.16	0.157
$\frac{Var(z)}{mean(z)}$	15.31	15.05
# of zeros in dataset (Z=0)	38	38

Table 1: This is a comparison between our calculations and the calculations produced in the Lloyd-Smith paper. We were able to recreate the results using the Singapore SARs dataset and mimic the different distributions on top of the secondary cases.

In the Lloyd-Smith paper, the authors provide a plot of the expected proportion of transmission due to the proportion of infectious cases.

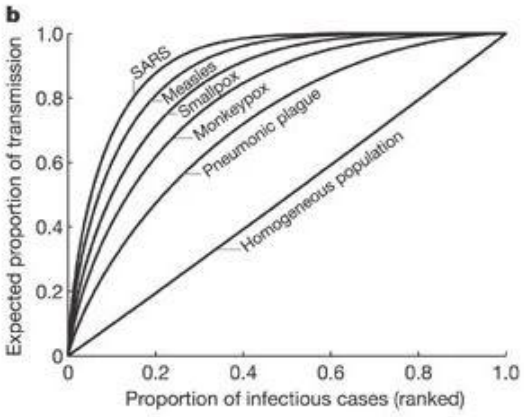


Figure 6: This plot originates from the Lloyd-Smith paper. It shows the different amounts of individual variation with SARS, measles, smallpox, monkeypox, and the pneumonic plague.

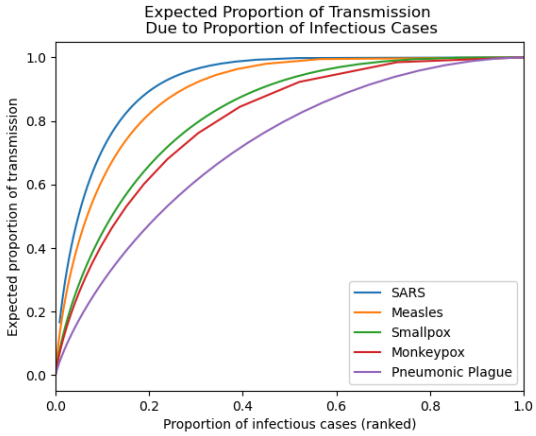


Figure 7: This plot depicts the expected proportion of transmission due to the proportion of infectious cases. We were able to replicate the Lloyd-Smith plot fairly accurately.

To recreate this plot, we used the R_0 and k values from the Supplementary Table 1 in the Lloyd-Smith paper. First, we found the $F_v(x)$, the cumulative distribution function of the gamma distribution with a shape parameter of k and a scale parameter of $\frac{R_0}{k}$. Next, we calculated $F_{trans}(x)$, the expected proportion of transmission with an individual reproductive number $v < x$. The equation for this was given in the Supplementary Notes of the Lloyd-Smith paper: $\frac{1}{R_0}\int_0^x u f_v(u)du$. This equation required integrating the probability density function of the gamma distribution with the same parameters as before. To find the expected proportion of transmission with $v > x$, $1\text{-}F_{trans}(x)$ was calculated. For the graph, $1\text{-}F_v(x)$ was plotted on the x-axis and $1\text{-}F_{trans}(x)$ was plotted on the y-axis.

3.1 Effects of Superspreading

Along with reproducing the same data from Lloyd-Smith et al., we took a look at some papers which talked about the variety of effects superspreading has on a population.

A paper from Ashish et al. in 2022 [4] decided to take a closer look at how superspreading events (SSEs) can cause variants of COVID. They found that when an SSE occurs and a single infection leads to twenty or more new infections, then that variant of the disease can survive as it bypasses the slow beginning of the exponential phase and moves into the rapid growth phase. If a variant of COVID or some other infection has an SSE early on in its lifetime, then the chances are greater that it will rise to become a dominant variant. Otherwise, more often than not, it will end up going extinct without one of these initial SSEs. In order to prevent dangerous variants of diseases from emerging, it is important to prevent superspreading events as best as we can. To perform their analysis, the researchers used part of the model from the Lloyd-Smith paper along with information about the probability of infection based on viral load. Using this information, they

were able to determine how many people need to become infected before a variant of a disease is dominating others.

Another interesting paper, by Metzger et al., brought up the use of superspreading to help promote the infection of people with therapeutic interfering particles (TIPs) [7]. We often hear about how superspreading events are dangerous, but this paper takes a look in the opposite direction by using superspreading as a means to help those who have HIV. The idea behind TIPs is that it lacks the ability to self replicate but retains HIV's genome packaging signal, therefore parasitizing off of HIV's movement. When HIV reproduces and/or moves, so will TIPs. The researchers modeled how this type of therapy may move within a population and what effects it may have. They found TIPs will typically infect high risk patients even if a small portion of the population is first injected with TIPs. Over time, it will spread quickly by following along SSEs. From here, it was found that if this therapy has the same effect as a few other therapies tested on HIV, then it could result in a decrease in prevalence from 30% to 6.5% after around 30 years. Thus showing that if it were possible to make a therapy that follows the spreading pattern of a disease, then it would be quite effective at nullifying it. Though of course, there are ethical implications of enacting such a plan as not everyone would agree to take a vaccine or be forced to have a certain therapy spread to them without their consent. Therefore, technologies like this one will need more improvements before it could be used in the real world.

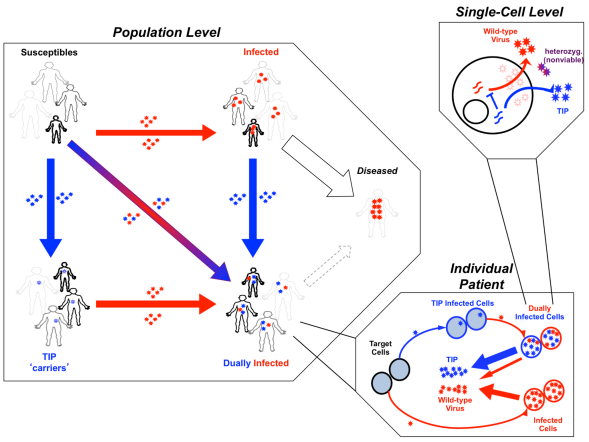


Figure 8: From the Metzger et al. paper, showing an example of what their TIPs spreading around looks like [7].

Another interesting study from Fukui et al. brings up the use of a different probability style when referring to superspreading [3]. The researchers use the fat-tailed probability distribution to better model superspreading effects from COVID. Then, stacking this on top of an SIR model, they found that superspreading events will continue to occur even after there is a large infected population size. While most other papers say that SSEs disappear when the infected population is large due to the law of large numbers, this paper emphasizes how it is still important to be wary of superspreading events. Even after a large portion of the population has been infected, superspreading events can still occur and infect more people. The paper demonstrates that even with a large infected population, it is still important to have superspreading prevention methods in place.

4 Discussion

Overall, it seems that the foundation for modeling superspreading events comes from the use of different statistical distributions and drawing a random number based upon the distribution. From this random draw, we typically notice that most infections die out rapidly but a small number end

up causing a ridiculously high amount of infections.

It might seem like after a large number of infections is reached, it is not as important to track superspreading events, as the law of large numbers comes into play. One paper in the literature seemed to counteract this point by using the fat-tailed distribution for its number of draws [3]. It pointed out that even if the number of infections becomes high, there will still be superspreading events that occur, and these superspreading events can still cause a large increase in the total number of cases. This is shown both in their model and in real life, where a few superspreading events had popped up even with a high number of COVID cases already present.

Every source in the literature agrees that superspreading events are frightening as they can shoot up R_0 values quickly and can cause variants of concern to arise at much higher rates. This makes it important for communities to put in methods of preventing SSEs from occurring. Currently, the best methods towards preventing these events involve being mindful of how many people are close together, the state of the ventilation systems in buildings, and being aware of coughing, sneezing and other aerosoles being spread around. By being mindful of these practices, superspreading events can be limited, thus lowering the R_0 values for different diseases [1].

Looking forward into the future, it may be possible to use superspreading as a means for health as mentioned by Metzger et al [7]. However, the ethical implications of such devices need to be considered before being employed in the real world. In the future, models for superspreading will hopefully improve so that we have the means of better understanding what is happening in our communities after different infections sweep through them. Using this information, we will be able to counteract infections and prevent superspreading events from occurring as often as they do now.

5 References

[1] Althouse BM, Wenger EA, Miller JC, Scarpino SV, Allard A, et al. (2020) Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. PLOS Biology 18(11): e3000897. https://doi.org/10.1371/journal.pbio.3000897

[2] CDC. (2003). Severe acute respiratory syndrome --- Singapore, 2003. Centers for Disease Control and Prevention. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5218a1.htm

[3] Fukui, M., & Furukawa, C. (2020, June 26). Power laws in Superspreading events: Evidence from coronavirus outbreaks and implications for SIR models. medRxiv. https://www.medrxiv.org/content/10.1101/2020.06.11.20128058v2.full

[4] Goyal Ashish, Reeves Daniel B. and Schiffer Joshua T. 2022 Multi-scale modeling reveals that early super-spreader events are a likely contributor to novel variant predominance J. R. Soc. Interface.1920210811 http://doi.org/10.1098/rsif.2021.0811

[5] Granik, A. (2021, June 5). Fitting and visualizing a negative binomial distribution in python. Medium. https://anton-granik.medium.com/fitting-and-visualizing-a-negative-binomial-distribution-in-python-3cc27fbc7ecf

[6] Lloyd-Smith, J., Schreiber, S., Kopp, P. et al. Superspreading and the effect of individual variation on disease emergence. Nature 438, 355–359 (2005). https://doi.org/10.1038/nature04153

[7] Metzger VT, Lloyd-Smith JO, Weinberger LS (2011) Autonomous Targeting of Infectious Superspreaders Using Engineered Transmissible Therapies. PLOS Computational Biology 7(3): e1002015. https://doi.org/10.1371/journal.pcbi.1002015

6 Code

The code for all our results can be found in this GitHub repository.