Effect of Contact Tracing and Lockdown on COVID-19 in Karnataka

Siva Athreya* Nitya Gadhiwala†

July 16, 2020

Abstract

We study the number of secondary infections across eight clusters of COVID-19 in Karnataka during the period March 9th to June 26th, 2020. For each cluster, using a Negative Binomial model, we estimate the effective basic reproduction number R_{eff}, dispersion and provide a confidence interval for the latter. We find that R_{eff} is less than 1 indicating benefits of contact tracing and quarantine measures put in place. However with dispersion k being small is indicative of individual variation in secondary infections. One can calculate the probability of super-spreading events using the model. We also show that the effective basic reproduction for clusters has variation across age, time, and generations.

1 Introduction and Preliminaries

For COVID-19, in the absence of a vaccine, key measures to contain infection spread have been lockdowns, contact tracing, quarantine, testing along with wide publicity of social distancing norms, hygiene guidelines, awareness of the symptoms of the disease and treatment. There are many efforts to understand control measures such as lockdowns, contact tracing and quarantine with respect to COVID-19 spread using stochastic models. In [6], using a stochastic transmission model it has been concluded that highly effective contact tracing and case isolation is enough to control a new outbreak of COVID-19 within 3 months and that the probability of control decreases with long delays from symptom onset to isolation, fewer cases ascertained by contact tracing, and increasing transmission before symptoms. In [5] the authors seem to suggest that COVID-19 spreads too fast to be contained by manual contact tracing, but could be controlled by a contact-tracing app which is faster, more efficient, and on a larger scale. They claim that by targeting recommendations to only those at risk, epidemics could be contained without resorting to lockdowns.

Contact tracing and other controlled measures were also used by countries during the Severe Acute Respiratory Syndrome (SARS) epidemic. In [4], the authors use detailed epidemiological data from Singapore and epidemic curves from other settings, to estimate reproductive number for SARS in the absence of interventions and in the presence of control efforts. They conclude that that a single infectious case of SARS infects about three secondary cases in a population that has not yet instituted control measures. In [3], the authors study the first 10 weeks of the SARS epidemic

*8th Mile Mysore Road, Indian Statistical Institute, Bangalore 560059, India. Email: athreya@isibang.ac.in
†8th Mile Mysore Road, Indian Statistical Institute, Bangalore 560059, India. Email: bmat1826@isibang.ac.in

in Hong Kong. The epidemic was characterized by two large clusters—initiated by two separate "super-spread" events (SSEs)—and by ongoing community transmission. Using a stochastic model, they compute basic reproduction number and transmission rates and conclude that the result of reductions in population contact rates, improved hospital infection control and rapid hospital attendance by symptomatic individuals resulted in fall of transmission rate and decline of the epidemic.

In [2], they argue that using only the basic reproduction number can obscure the individual variation in infectiousness. Their motivation being 'super-spreading events' in which certain individuals infected unusually large numbers of secondary cases (5–10 in the SARS epidemic). They studied contact tracing data from eight directly transmitted diseases, and showed that the distribution of individual infectiousness around the basic reproduction number is skewed. Using various models they then proceed to compare effect of individual specific control measures versus population-wide measures. They conclude that super-spreading event are a normal feature of disease spread and give a formal definition of the same.

Since 9^{th} March 2020, the Government of Karnataka has been providing detailed media bulletins containing specific guidelines on the virus and information on each patient to have tested positive in the state. The bulletins on the tested positive patients contain information regarding how each one of them contracted the virus (either travel or by being a contact of someone who has known to have tested positive for COVID−19) or how they came to be tested (either as a Severe Acute Respiratory Infection patient or someone with Influenza like symptoms).

In this article we study the trace history provided in the media briefs and try to understand the spread of the disease in the period from 9^{th} March till 26^{th} June 2020 given the control measures taken by the state. Efforts by the State to reduce transmission have been manifold and were expected to have a substantial impact on reducing the size of the epidemic. From the trace history we were able to divide the patients who tested positive in to several clusters. For each cluster we find that effective basic reproduction number is less than 1 but variance is larger than the mean. The distribution of secondary infections across all clusters is very skewed at 0 due to the control measures taken. We fit a Negative Binomial model that showed significant dispersion across clusters. Thus though the clusters are going to die out under the controlled environment there is a reasonable chance of super-spreading events. We are able to predict their frequency within these clusters using this model. Further we show that the clusters have age, time and generational variation with regard to basic reproduction number. These can be explained by external factors that prevailed during this period for each cluster.

Remark 1. *A word of caution before we proceed. Our entire work is based on data provided in the Media Bulletins, [9]. These are dependent on the contact tracing procedures and testing policy followed by the government, unknown to us. From Figure 1, the fraction of tests positive is around 2% at this time and has been as high as* 4.09% *on April* 2^{nd}*, 2020 with a low at* 0.82% *on May* 14^{th}*, 2020. The number of total tests conducted up to* 26^{th} *June is* 544054 *with the test positive fraction being at* 2%*. These provide a comprehensive count of testing numbers in the state but not cluster wise testing data.*

Needless to say, the number of Infected in the population differs from the number of positive test results. So equating the number of those tested positives as number of infected individuals may be an error, because every individual in the population has not been tested. Thus for any inference or conclusion on true infection growth we must take into consideration the different policy/rates of testing, population density, contact tracing, quarantine measures, and biological aspects of this epidemic.

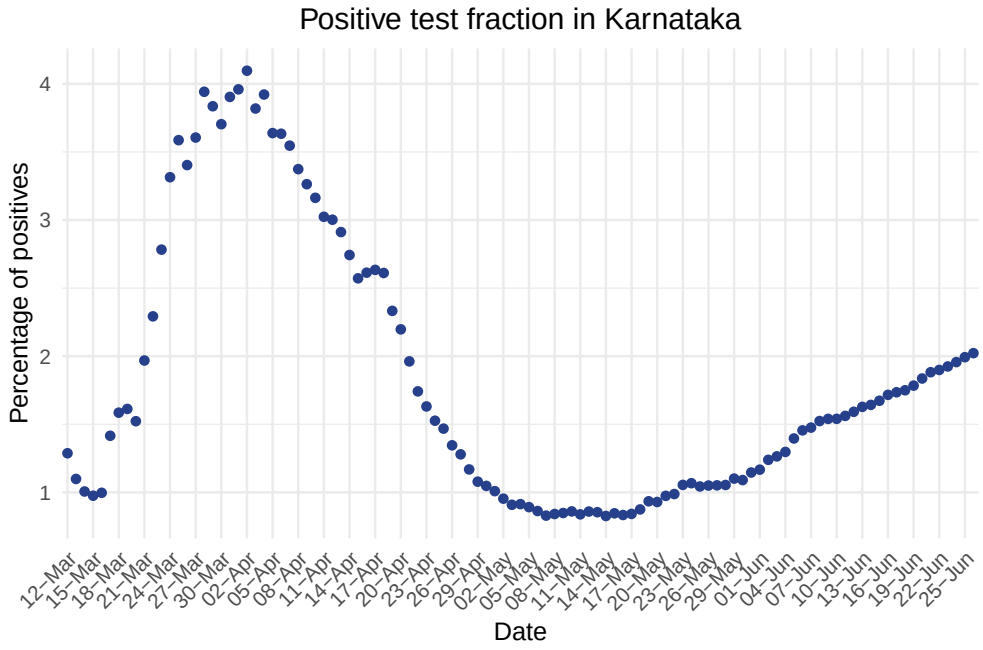


Figure 1: The above graph represents the test positive fraction in Karnataka. For each day, we plot on the y-axis the percentage of test positives in Karnataka up to that date (i.e., on each day, we plot the percentage of the total positives up to that day over the total number of tests conducted up to that day.)

Rest of the paper is organised as follows: In Section 2 we describe the clusters in Karnataka that we consider and calculate the effective reproduction number, R_{eff}, and the dispersion parameter,k, along with it's 95% confidence intervals. In Section 3, we present our observations and list out certain precautions before making inferences on the findings. In Section 4, we point to the data sources used and made available by us. The appendix contains in Section A–the model assumed by us, in Section B– the method of estimating parameters, in Section C – the χ^2 test used and in Section D– the method of used to estimate confidence intervals.

Acknowledgements: We would like to thank Gautam Menon for introducing us to the question and also for pointing us to [2]. Further we would like to thank P. Shankar, Rajesh Sundaresan, Deepayan Sarkar, Mohan Delampady, and Abhiti Mishra for useful discussions.

2 Estimating R_{eff} and Dispersion across infected clusters.

In the daily media briefs provide by the Karnataka State Ministry of Health and Family Welfare, apart from issuing regular guidelines for the public, contain detailed information of screening of air/sea passengers, people under observation, number of positive tests on each day and details of each infected patient [including: age, sex, district, source of infection, etc.] We first divide up the cases into clusters from their place of origin, for example "From Europe" or "Pharmaceutical Company in Nanjangud, Mysore". Then in each cluster we place all the patients who got it independently from the place of origin and then recursively add the patients who they passed the infection to.

We divide the entire set of COVID-19 tested positive patients in Karnataka into the following major clusters.

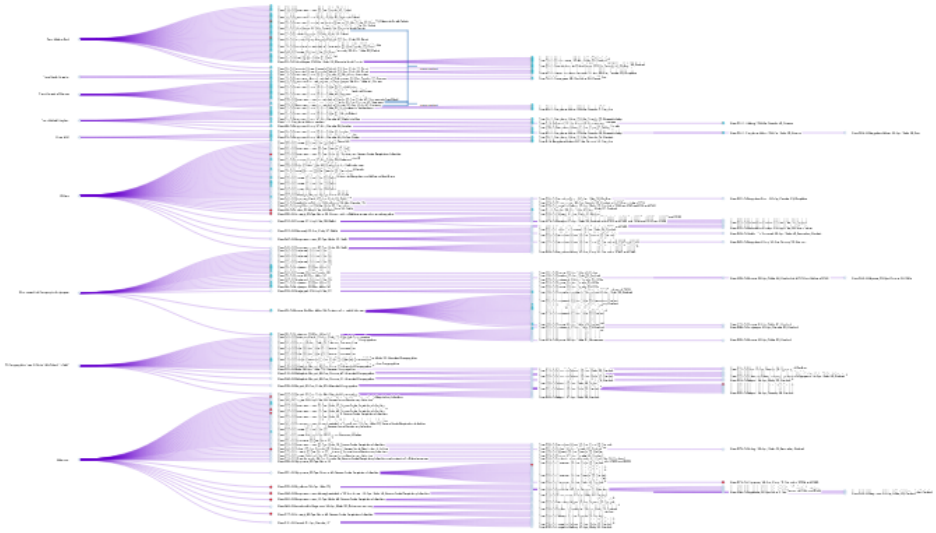


Figure 2: Karnataka Trace History

- Imported
  - From Middle East
  - From USA
  - From United Kingdom
  - From Rest of Europe
- Migration
  - From Maharashtra
  - From Rajasthan
  - From Southern States
- Local
  - Influenza like illness (ILI)
  - Severe Acute Respiratory Infections (SARI)
  - Unknown
  - Others
  - Pharmaceutical Company-Nanjangud
  - T.J. Congregation in Delhi
  - Containment Zones

Table 1: Clusters of COVID-19 in Karnataka from 9th March - 26th June 2020

In the interactive graph, on our accompanying website, we plot the trace history as a tree like graph. The first generation nodes [at depth one] in the graph of each cluster are the patients who got the infection directly from the place of origin of the infection. These patients we shall call as 'parents' of the cluster. The 'children' are the people who contracted the disease from the people labelled as 'parents', that is, they are at a depth of two in the trace history chart. Similarly, 'grandchildren' and 'great grandchildren' have depth three and four respectively. A part of the graph can be seen in Figure 2.

One thumb rule for disease spread, including COVID-19 anecdotally, is the 20/80 rule. The rule states that 80% of the secondary infections arise from 20% of the primary infections. As seen in Figure 3 if we look at patients in Karnataka who tested positive on or before 3rd May and their descendants then the rule holds true. Figure 3 attempts to demonstrate the heterogeneity of the infectiousness of the individuals infected with COVID-19 in Karnataka. It can be observed that for Karnataka, almost 20% of the individuals with the highest infectiousness are responsible for 70% of the total infections. In the case of a perfectly homogeneous population of infected individuals

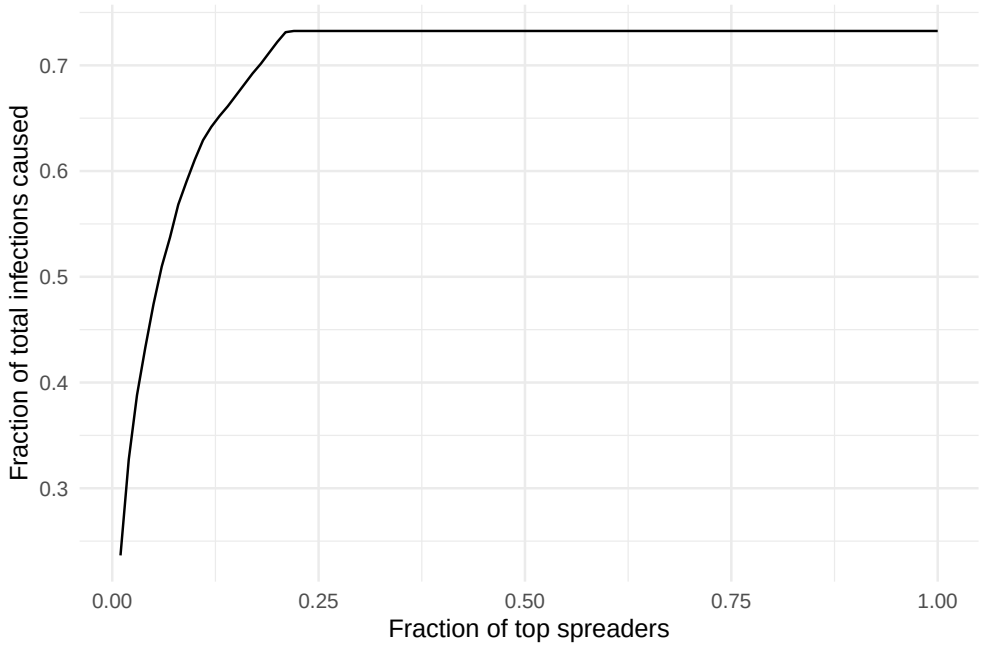


Figure 3: The plot considers those individuals who were infected before 3-May along with all those cases that can be traced back as contacts of them. We rank the infected individuals in terms of the number of secondary infections caused by them. Then we consider the top x fraction of them. And plot x on the x-axis and the fraction of the total infections infected by them on the y-axis.

(i.e., every infected individual infects equal number of healthy individuals), would represent the $x = y$ line. The large deviation from the $x = y$ straight line represents the heterogeneity in the infected individual population.

In epidemiology, the "basic reproduction number", denoted by R_0, of an infection can be thought of as the expected number of cases to have contracted the infection directly from one case. Thus on an average each infected person passes on the infection to R_0 many healthy individuals. The number R_0 is by no means a unique number for a disease. It greatly varies with time (start versus end), variation in a region's population density and depends heavily on interventions put in place to curb the spread of the infection.

In Karnataka, from the very beginning quarantine measures and contact tracing were put in place for all the imported clusters. To contain the spread of COVID−19 infections in India the union government started a strict lockdown on 25^{th} March and relaxed them over 5 phases as follows: Lockdown Phase 1 (25^{th} March – 14^{th} April) and Lockdown Phase 2 (15^{th} April – 3^{rd} May) were the strictest in terms of mobility; Lockdown Phase 3 (4^{th} May – 17^{th} May) and Lockdown Phase 4 (18^{th} May – 31^{st} May) included relaxations of travel between states; and Unlock 1.0 (1^{st}–30^{th} June), Unlock 2.0 (1^{st}–31^{st} July) which have had considerable relaxations. Thus in Karnataka during the period 9^{th} March - 26^{th} June we are observing the COVID-19 infection spread in a controlled environment. So whenever we calculate basic reproduction numbers we are actually calculating effective reproduction number of the disease during this period.

We begin by considering entire tested positive population in the state and examine the distribution of number of new infections designated as contacts of earlier infections (See Figure 4).

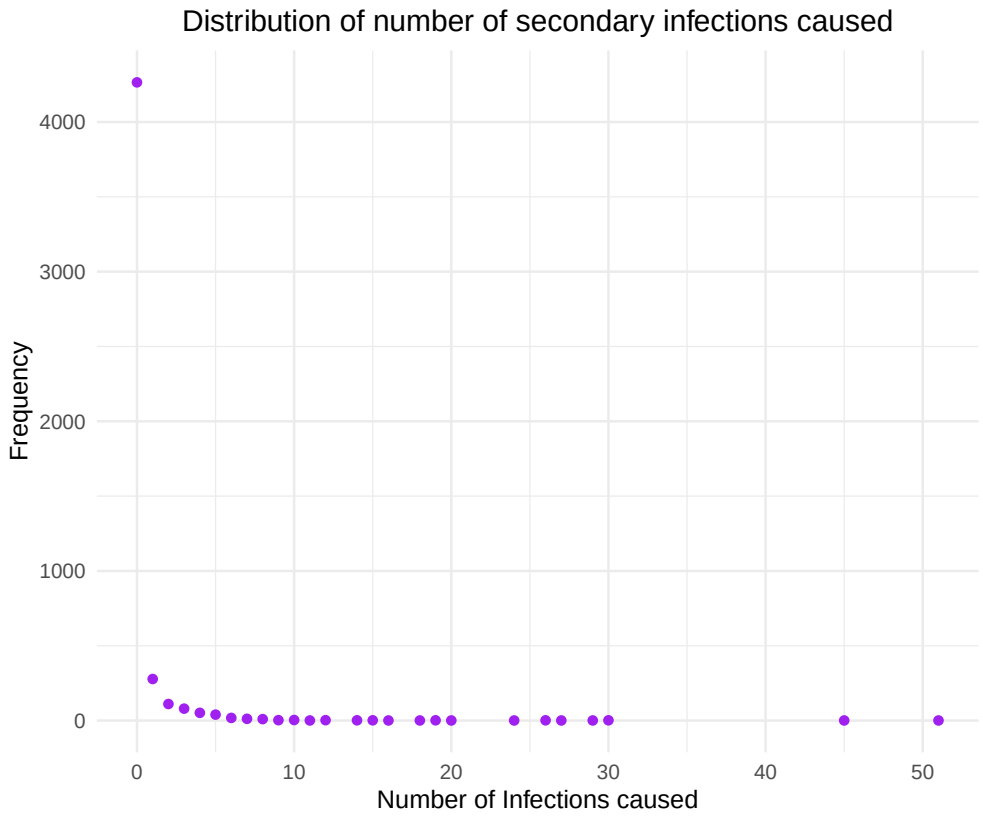


Figure 4: The above scatter plot considers the COVID−19 patients in Karnataka and represents the distribution of the number of infections caused by each patient. The patients belonging to the 8 clusters in study are considered here. We plot the frequency distribution of the number of infections assigned to each infected individual as their contacts. The plot above has the number of infections caused on the x-axis and the number of patients that have caused x many infections on the y-axis.

In Figure 4, a large peak is seen at 0 infections caused. It can be seen that only 9 individuals in the population of 4895 have passed the infection on to more than 20 people. This could be the result of a super-spreader phenomenon or perhaps an effect of how the contact tracing and testing is performed. Assigning them as definitely arising from one particular individual will need a more careful understanding of the latter. One can further note that due to effective quarantine measures there are 4265 infected individuals who have not passed the infection on to anyone else.

We will study eight clusters from Table 1. Namely,

- From Southern States
- Influenza like illness
- Severe Acute Respiratory Infections
- Containment Zones
- Unknown
- Others
- Pharmaceutical Company- Nanjangud
- TJ Congregation in Delhi

The clusters we are studying have all begun before 3$^{\text{rd}}$ May 2020 and have more than 50 individuals in total. A couple of them, Pharmaceutical Company-Nanjangud and TJ Congregation in Delhi have not had any new test positive cases in the last one month. These two clusters also contain no active cases as of today. The other clusters are still growing and all calculations and analysis

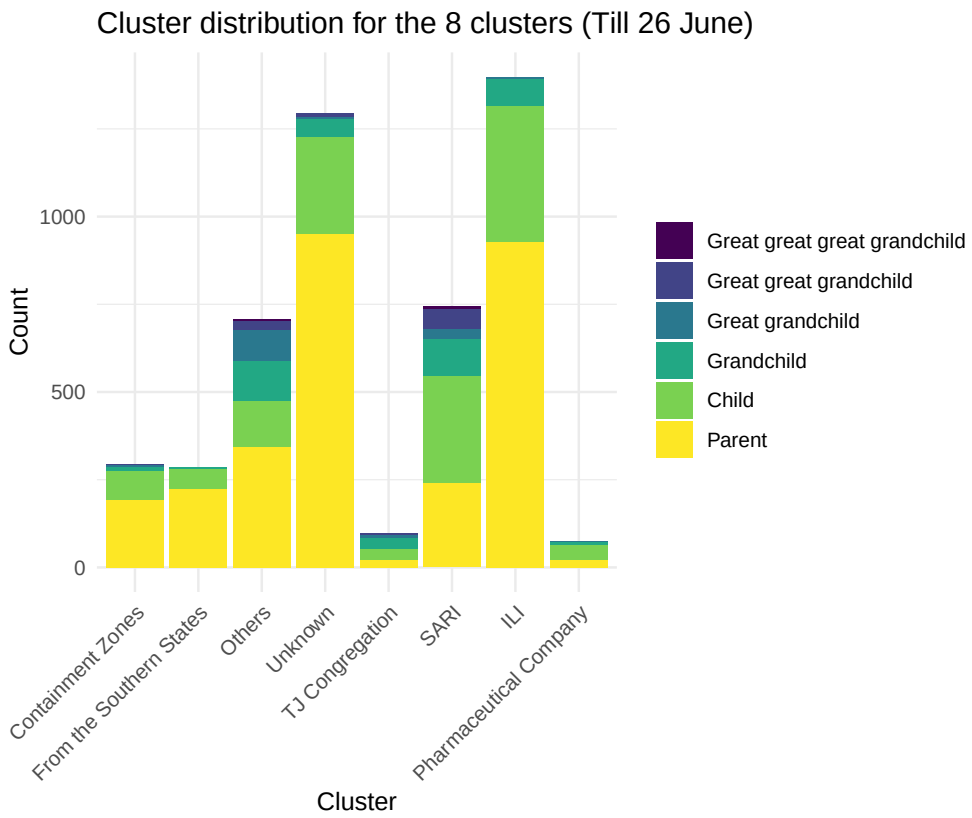


Figure 5: The above plot is a stacked histogram displaying the distribution of the 8 clusters we consider till 26-June. The histogram represents the number of infections that belong to each of these clusters and each bar has been further filled to denote the number of primary infections, secondary infections and so on.

for them is dated to 26^{th} June, 2020 in this note. We have omitted two clusters which satisfy this criteria from analysis, namely: From Maharashtra and From Middle East. We shall explain the reasons below. We present a summary distribution of parents, children, grandchildren, and great grandchildren in each of the 8 clusters we consider in Figure 5.

We now focus on the distribution of children for each of the clusters. For each individual i in the cluster we will denote the number of children (or the number of tested positive cases) assigned to patient i by y_i. The mean of y_i is the basic reproduction number R_{eff}. In Table 2 we present a comparison of the summary distributions across clusters.

From Table 2 we see that the variance does not match the mean. Further, as noted in Figure 3 heterogeneity in the infectiousness of each individual implies that solely R_{eff} isn't a good measure of the infection spread. To account for both of these we now consider the standard method of mixture of Poisson distributions to model the data set. For each cluster, using the Negative Binomial with mean R_{eff} and dispersion k (See [2] and Section A for details) as the offspring distribution, we will use the Maximum Likelihood method for estimating R_{eff} and k (See Section B for details).

Further one also notes that in Table 2, the Maximum column is not small. This might be indicative of a super-spreading phenomenon. A general protocol for defining a Super-spreading event was given in [2]: (1) estimate the effective reproductive number, R_{eff}, for the disease and population in question; (2) construct a Poisson distribution with mean R_{eff}, representing the expected range of Z (without individual variation); (3) define an Super-spreading event as any infected individual who infects more than Z_n others, where Z_n is the n^{th} percentile of the Poisson(R_{eff}) distribution.

Cluster	Size	Zeros	Maximum	R_{eff}	Variance
Unknown	1295	1173	27	0.2625	1.637
Pharmaceutical Company	73	53	24	0.726	8.757
From the Southern States	286	257	7	0.2168	0.6897
Others	707	593	51	0.5191	6.502
TJ Congregation	97	70	15	0.7732	3.823
SARI	746	622	45	0.6743	8.521
ILI	1398	1243	30	0.3369	2.355
Containment Zones	293	254	7	0.3447	1.199

Table 2: For the 8 clusters we consider, the above table contains information on the following. The size column represents the number of infected individuals belonging to each cluster. The Zeroes column denotes the number of patients who haven't been responsible for any new infections as their contacts. The Maximum column represents the highest number of individuals assigned as contacts of a particular individual for that cluster. The R_{eff} column represents the mean of the number of infections assigned to each existing infection as their contacts. The Variance column represents the variance of the same.

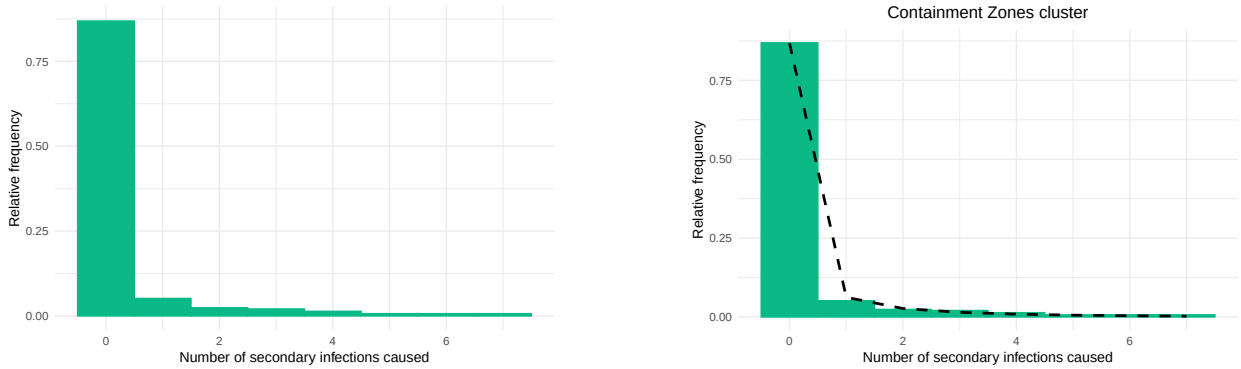
A 95^{th} or 99^{th}-percentile super-spreading event is any case causing more infections than would occur in 95% or 99% of infectious histories in a homogeneous population. If R_{eff} and k have been estimated then one can use the definition and the Negative Binomial model to understand the probability with which such events will occur. We will try to understand this phenomenon with regard to the 8 clusters under observation.We begin by explaining in detail the analysis for the Containtment zone cluster and then present a summary for the remaining clusters.

In late April, as the number of cases in Karnataka reached around 500, certain areas of the state were demarcated as Containment Zones. Areas in such zones, as per government directives, had restricted access (both into and out of) and featured quarantine for the residents in them. The first coronavirus case assigned to the Containment Zones cluster tested positive on 8^{th} May. The reason for their testing positive was listed as Contact of Ward 135 of the Containment Zone in Bangalore (Padarayanapura). Since then, count has reached 293 odd (see Table 2). The individuals in this cluster are either direct contacts of a Containment Zone or can be traced back as contacts of such individuals. The size of the cluster is 293, the maximum number of secondary infections assigned to an infected person is 7, there are 254 infected individuals in this cluster who have been assigned 0 secondary infections. The Maximum likelihood estimates for R_{eff} and k are given by

$$R_{\text{eff}} = 0.3447 \qquad \text{and} \qquad k = 0.09345.$$

The 95% confidence interval for k is $(0.06736, 0.15250)$ and the p-value from the χ^2-goodness of fit test is given by 0.5966. As the confidence interval is fairly small and the p-value is not less than 0.05 the Negative Binomial model can be selected for the data set.

If we were to consider a 99^{th} percentile event with the above $R_{\text{eff}} = 0.3447$ then an event causing more than 2 secondary infections would be considered a super-spreading event. In the Containment zone cluster there is a person who has been assigned 7 secondary infections, this would be considered a super-spreading event. Under the Negative Binomial model the probability of observing 7 secondary infections is 0.0027. This may indicate one of two possibilities, either a rare event has occurred or it is an effect of how testing policy and contact tracing are being followed. We shall discuss this further in the next section. In Table 4, we have computed the Maximum Likelibood estimators for R_{eff} and k for each cluster and also performed the χ^2-goodness of fit test (see Section C for details regarding the goodness of fit).



The histogram on the left indicates the number of secondary infections assigned to infected persons in Containment Zones via contact tracing in Media briefs. On the right, the dashed line represents the the Negative Binomial probability mass function with mean R_{eff} and dispersion k as calculated above.

Cluster	Size	Maximum	R_{eff}	Variance	k	p-value
Containment Zones	293	7	0.3447	1.199	0.09345	0.5966
ILI	1398	30	0.3369	2.355	0.06428	0.736
SARI	746	45	0.6743	8.521	0.08023	0.4698
TJ Congregation	97	15	0.7732	3.823	0.2138	0.1138
Others	707	51	0.5191	6.502	0.09214	0.8318
From the Southern States	286	7	0.2168	0.6897	0.08424	0.5409
Pharmaceutical Company	73	24	0.726	8.757	0.1839	0.002671
Unknown	1295	27	0.2625	1.637	0.05792	0.1225

Table 4: The above table is an extension of Table 2. The above table, apart from some of the columns in Table 2 contains the dispersion parameter, k and the p-value of the χ^2 goodness of fit test. (See Sections B and C for details.)

The R_{eff} and k calculated here are those after many restrictions and hence are in no way representative of the R_{eff} and k of the disease. Strict contact tracing and isolation have increased the level of control and hence the R_{eff} calculated here is the effective R_{eff} and k, and may be different from that calculated in other works. Even with the most stringent contact tracing policies, a few contacts are bound to be left out in which case the effective R_{eff} may be larger than the value calculated here.

The p-values in the last column are not small for all clusters except for the cluster at Pharmaceutical Company at Nanjangud. This cluster has a very high variation, a maximum data point at 24 (i.e. one person who has been assigned to 24 secondary infections) and also a significant proportion at 1 infection caused. One can also see that the confidence interval for the dispersion is quiet large as well as seen in Figure 6. In Figure 7 we present, for each cluster, the histogram of the number of infections caused along with plot of the Negative Binomial fits and in Figure 8, the histogram of the Negative Binomial probabilities for each cluster along with their 95% and 99% percentiles.

In Table 5 we have computed 90^{th}, 95^{th} and 99^{th} percentile of the Poisson distribution with the respective R_{eff} for each cluster. By our earlier stated protocol these events define the Super-spreading events. In Figure 8 we have plotted the histogram of Negative Binomial probabilities for each cluster assuming the parameters estimated using the maximum likelihood estimate. We have marked the 95^{th} and 99^{th} percentile for these distributions in the plot. Finally, Table 6 and Figure 6 we provide the confidence intervals for the 8 clusters.

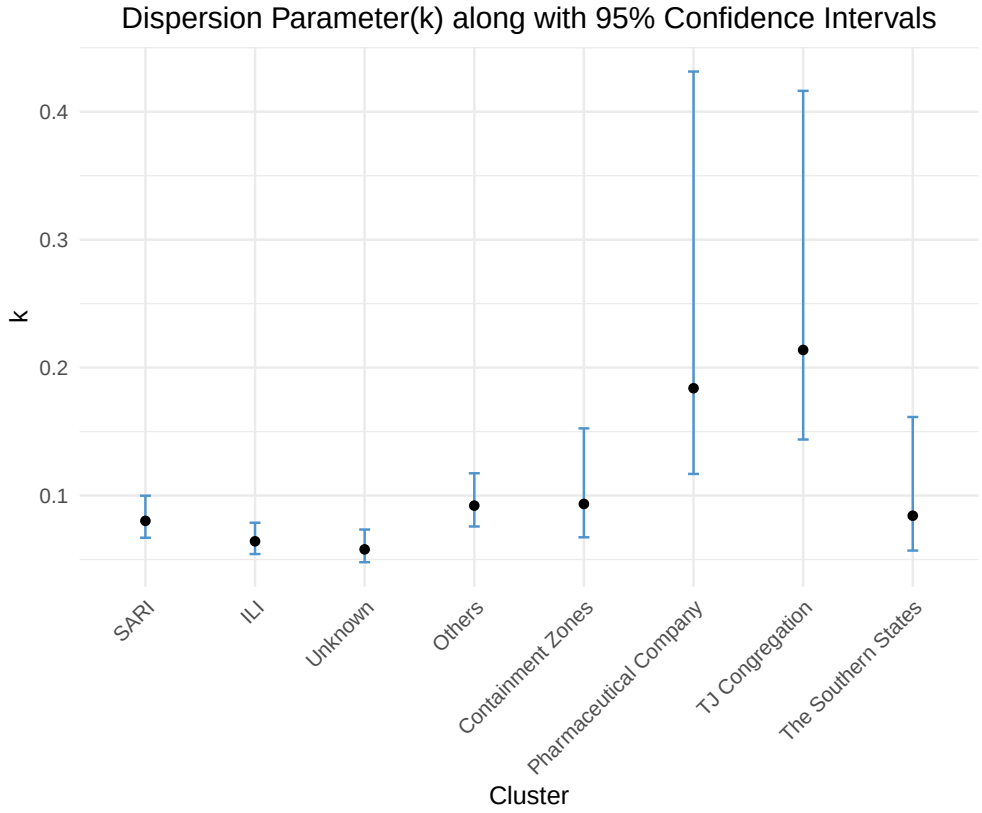


Figure 6: The above plot contains the calculated values of the dispersion parameter, k, along with the 95% confidence intervals for the 8 clusters in observation.

Cluster	R_{eff}	90th percentile	95-percentile	99-percentile
Containment Zones	0.3447	1	1	2
ILI	0.3369	1	1	2
SARI	0.6743	2	2	3
TJ Congregation	0.7732	2	2	3
Others	0.5191	1	2	3
From the Southern States	0.2168	1	1	2
Pharmaceutical Company	0.726	2	2	3
Unknown	0.2625	1	1	2

Table 5: The above table assumes a Poisson distribution for the number of infections caused by each individuals. Having calculated R_{eff} as the sample mean, the above table contains the 90^{th}, 95^{th} and 99^{th} quantiles of the Poisson(R_{eff}) distribution. If one assumes a Poisson distribution, the cases causing more than the above quantiles may be considered as super spreading events.

Cluster	left	k	right
SARI	0.06704	0.08023	0.09989
ILI	0.0543	0.06428	0.07874
Unknown	0.04788	0.05797	0.07344
Others	0.07582	0.09214	0.1174
Containment Zones	0.06737	0.09345	0.1525
Pharmaceutical Company	0.1169	0.1839	0.4314
TJ Congregation	0.1438	0.2138	0.4163
The Southern States	0.057	0.08424	0.1614

Table 6: The above table provides 95% confidence intervals for k. See Section D for methodology used to calculate them.

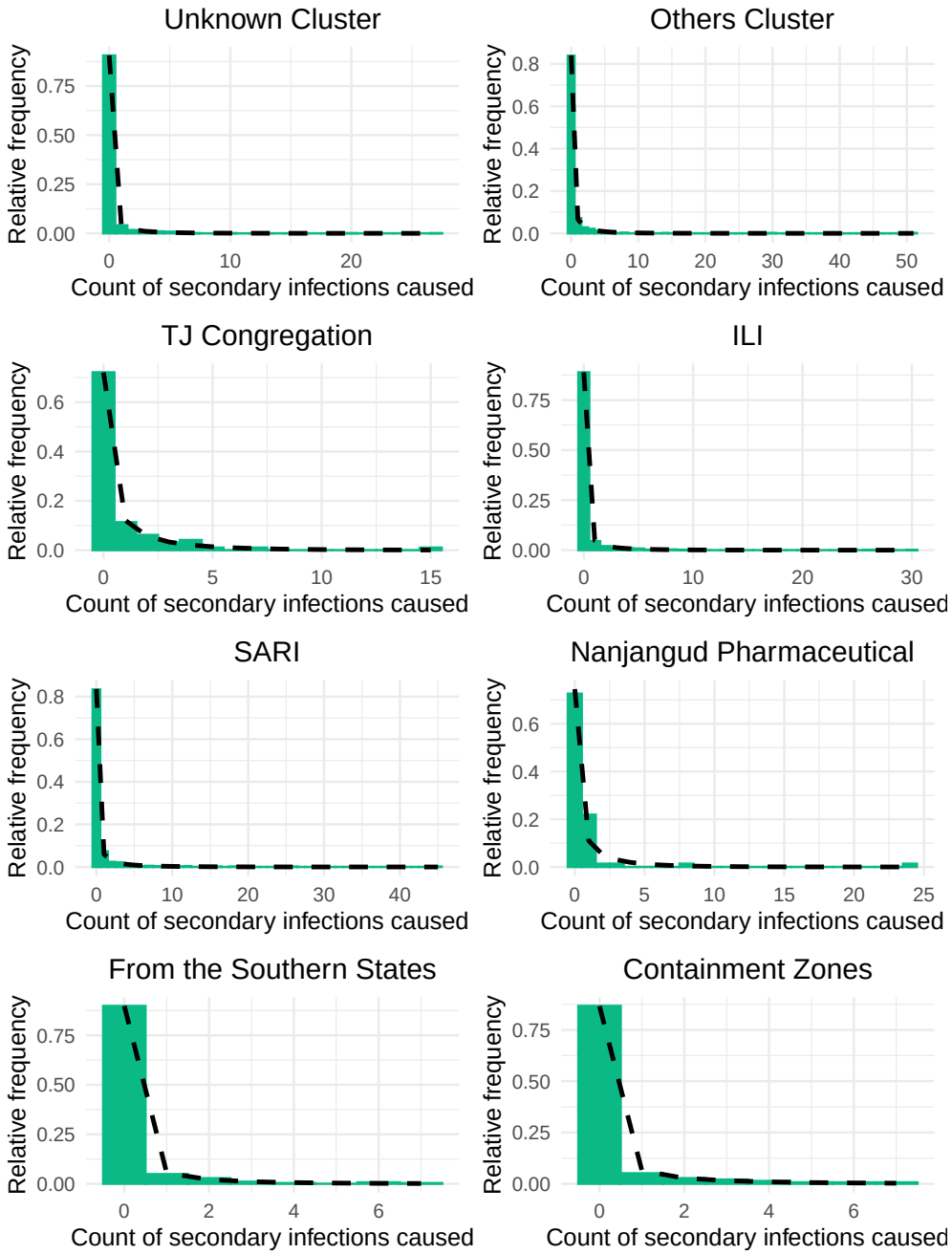


Figure 7: For the 8 clusters described, we plot the number of histogram of the number of secondary infections caused by each cluster along with the Negative Binomial fit. The green histogram is the offspring distribution and the black dotted line is the calculated Negative Binomial distribution assuming the offspring distribution. The negative binomial fit is calculated using the Maximum Likelihood Estimate.

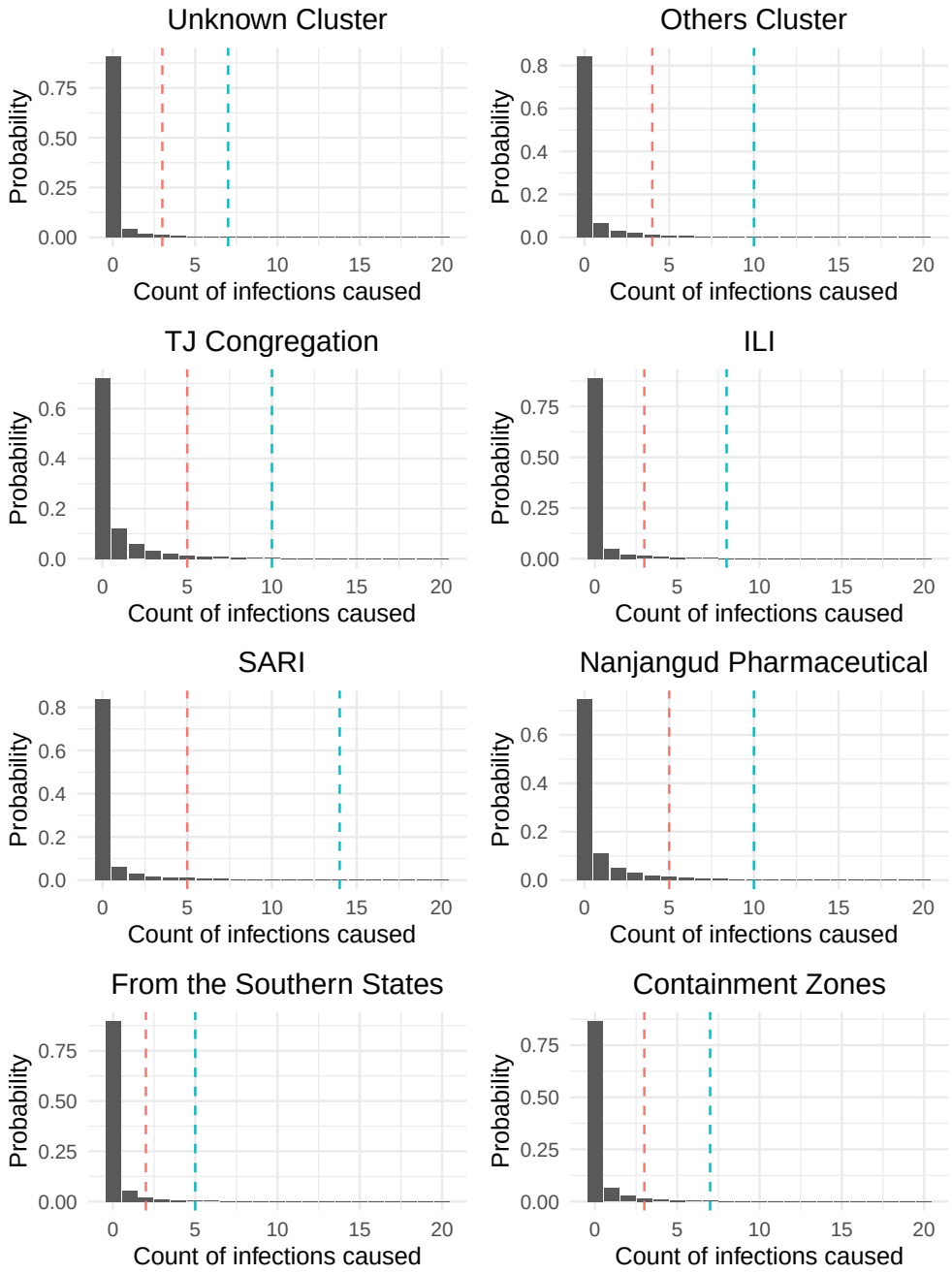


Figure 8: The above histogram is probability mass function of the Negative Binomial models for each cluster along with 95^{th} and 99^{th} quantile marked in red and blue respectively. One can use this to compute the probability of super-spreading events.

Cluster	Size	Maximum	R_0	Variance	k	p_val
From Maharashtra	5486	30	0.03992	0.3804	0.007059	0.5942
From Middle East	409	6	0.02934	0.1168	0.007965	0.3165

Table 7: For the two clusters From Maharashtra and From Middle East an analysis similar to that of Table 4 has been done.

We conclude this section with the two clusters that we had omitted from the detailed analysis above. The "From Maharashtra" cluster due to migration from Maharashtra. During Phase-3 of the lockdown, this cluster saw the most growth and dominated the test positive counts by a significant margin. The "From Middle East" cluster seems to have two phases. The first, before the lockdown was enforced during which international travel were suspended. The second, more recent, due to the repatriation flights from the region. We provide the Maximum Likelibood estimators for R_{eff} and k, along with the summary of them in Table 7. However we believe that we need to observe them for some more time to make any inferences.

3 Observations

The contact tracing for the eight clusters show that the offspring distribution is highly skewed. Thus we used a model as proposed in [2] to account for this variation. From Table 7 we have observed that $R_{\text{eff}} < 1$ for all the clusters. This will imply that all infection growth in all the clusters will eventually die out, regardless of the dispersion.

Generations within Clusters: The entire trace history is a measure of how contact tracing is being done and how infected individuals are being identified for testing. Thus, it need not necessarily capture the entire spread of the infection. Many of the contacts tested are contacts of multiple COVID-19 patients and assigning them to any one contact may not be indicative of the way the infection was passed on. It is also important to note that the parent to child relationship in trace history is indicative of the testing policy and contact tracing that was followed and need not be that of the infection spread.

The SARI cluster contains those patients who have a history of Severe Acute Respiratory Infection, and those who can be traced back as contacts of such patients. It should be noted that only the first generation of the patients in this cluster are those with a history SARI and the rest need not necessarily display similar symptoms. Hence, any inferences made on the SARI cluster isn't one on patients who have tested positive for COVID-19 and have a history of SARI. The same is true for the ILI cluster where the first generation contains those infected individuals with influenza like symptoms. For both these clusters, there is no information on the source of infection by the parents. It is likely that many were tested for COVID-19 due to testing policy that requires SARI patients to be tested. Thus contact tracing post testing positive for such patients may not necessarily indicate a parent-child relationship as far as the infection spread is concerned. However it is known that SARI patients have a high viral load of the infection. This can be seen in one of two ways. We have already seen in Table 5 that the 90^{th} percentile-super-spreading events are those larger than or equal to 2 and Figure 8 that the 95^{th}-percentile falls at 5 for the SARI cluster. Secondly, in Figure 9 where the mean number of offsprings of parents is higher than 1.5 though the R_{eff} for the cluster is less than 1.

Table 8 contains information on the 8 clusters in observation considering the generations separately. The maximum number of infections caused by an individual in the first generation is 30. An

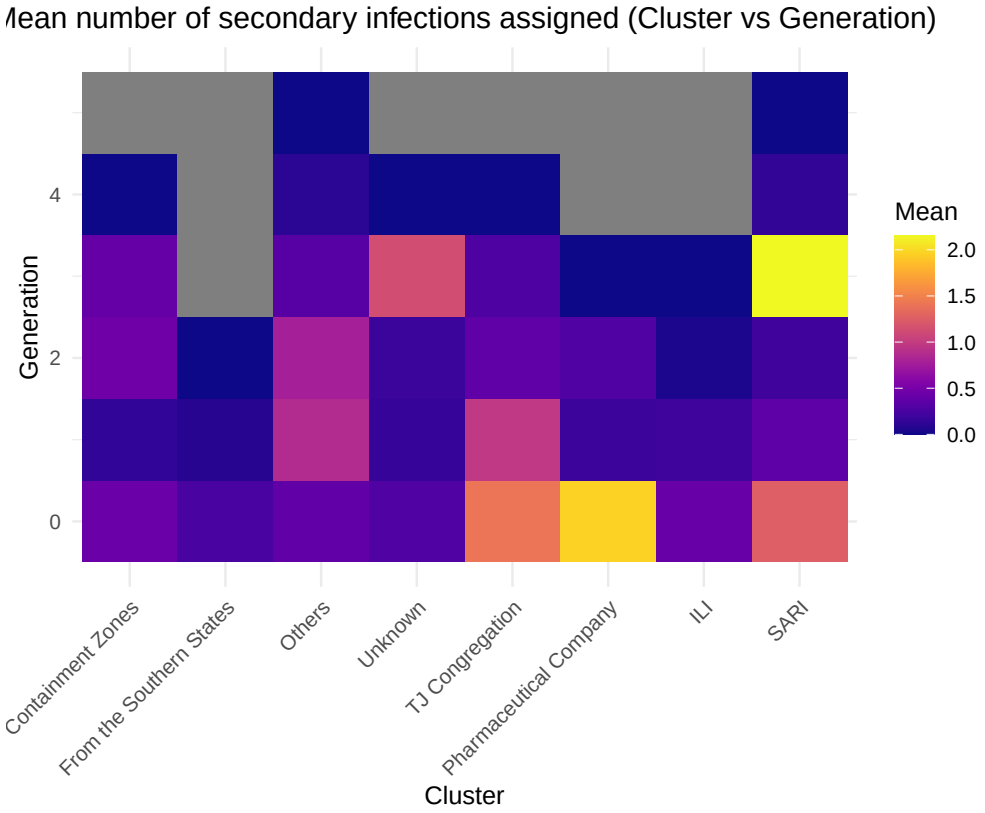


Figure 9: (Clusters across generations) Generation 0 refers to those patients who were the first in their respective clusters to be identified as infected. They may have had either some form of travel history or displayed certain symptoms. Generation $x+1$ represents the individuals who were infected by Generation x. The color of each tile represents the mean number of infections caused by each individual belonging to a particular cluster and generation.

individual in the Influenza like illness cluster and another in the Others cluster have caused 30 secondary infections each. Among the individuals in the second cluster the one to have caused 51 infections belongs to the Others cluster in Bellary. This patient was one of the first ones to be infected in a Steel plant in Bellary and hence may have had several infections assigned to him. It is observed that among those patients belonging to the Generation-4 (G Grandchildren), the mean, 0.7042, is significantly higher than the remaining generations. This is because of the small size of the cluster (142) and one of the patients being assigned as the contact of 45 other patients. While the highest generation that can be observed is Generation-6 (Great great great grandchild), they haven't been included in Table 8 as there isn't a Generation-7 and the individuals all have 0 infections assigned to them.

Cluster	Size	Zero	Maximum	Mean
Parents	2925	2528	30	0.4499
Children	1314	1176	51	0.3067
Grandchildren	403	344	14	0.3524
G Grandchildren	142	112	45	0.7042
GG Grandchildren	100	94	5	0.11

Table 8: The above table considers the different generations of infections as seen in Karnataka for the 8 clusters. For each generation the table contains the number of individuals in each generation, the number of patients causing zero secondary infections, the maximum number of infections caused by an individual in that generation and the mean number of infections caused by an individual in that generation.

Clusters across Dates: Before Phase 1 (25^{th} March - 14^{th} April) of the lockdown began, almost all the COVID−19 cases that were confirmed in Karnataka were either individuals who had some form of international travel history or those who were contacts of such individuals. These initial infections were very well contained as many of the infected individuals were confirmed and isolated quite early. Most of these clusters showed very few generations and very few infections caused by each of these individuals.

Both Phase 1 and Phase 2 (15^{th} April - 3^{rd} May) of the lockdown in Karnataka saw heavy restrictions on travel and nearly all services and factories were suspended. During this period a Pharmaceutical company in Nanjangud, Mysore, saw a sudden increase in the COVID−19 cases. Although the exact reason for the infection to reach the company is unknown, the first patient to be infected came in contact with health care workers treating COVID−19 patients. The first patient, a 35 year old male, was confirmed to be infected on 26^{th} March. Another local cluster to form was the TJ Congregation, which contained those who attended the TJ Congregation from 13^{th} to 18^{th} March in Delhi. The first patient to be confirmed in this cluster was on 2^{th} April. Both these clusters were very well contained and the last patients to be attributed to these clusters was confirmed on 29^{th} April and 21^{st} May respectively.

Another cluster which began during Phase 1 of the lockdown was the Severe Acute Respiratory Infection (SARI) cluster. In a change in testing policy by the government, all the patients with a history of SARI showing symptoms were tested. The first infection to be attributed to this cluster was on 7^{th} April.

An initiative taken by the government was to create Containment Zones in regions with many cases. The first case reported to have been in contact with a containment zone was on 24^{th} April. Since then this cluster saw a large fraction of the increase to occur during Phase 3 (4^{th} May - 17^{th} May) and Phase 4 (18^{th} May - 31^{st} May) of the lockdown. Phase 3 and 4 of the lockdown loosened restrictions on Domestic Travel and many infected individuals had some domestic travel

history. The state saw a large influx of infected individuals from states like Maharashtra, Gujarat, Tamil Nadu, Telengana and Andhra Pradesh. The clusters Others, From Maharashtra and From the Southern States saw large increases during this period. Among these, the highest number, by far, was that of individuals with travel history to Maharashtra.

The Phase 4 of the lockdown and the unlockdowns that followed saw a surge in Influenza like illness (ILI) and Unknown cases. The number of infected individuals coming from Maharashtra significantly decreased during the Unlock 1.0 (1^{st} June - 30^{th} June) and Unlock 2.0 (1^{st} July - 31^{st} July) phases. Since testing strategies are unknown to us, no immediate conclusion can be drawn. During these phases, the From Middle East cluster also saw a second wave of infections as several patients with international travel were observed. Despite this surge, this cluster was very well contained as most of the individuals caused very few secondary infections. This can be seen as the R_{eff} for these clusters is very low.

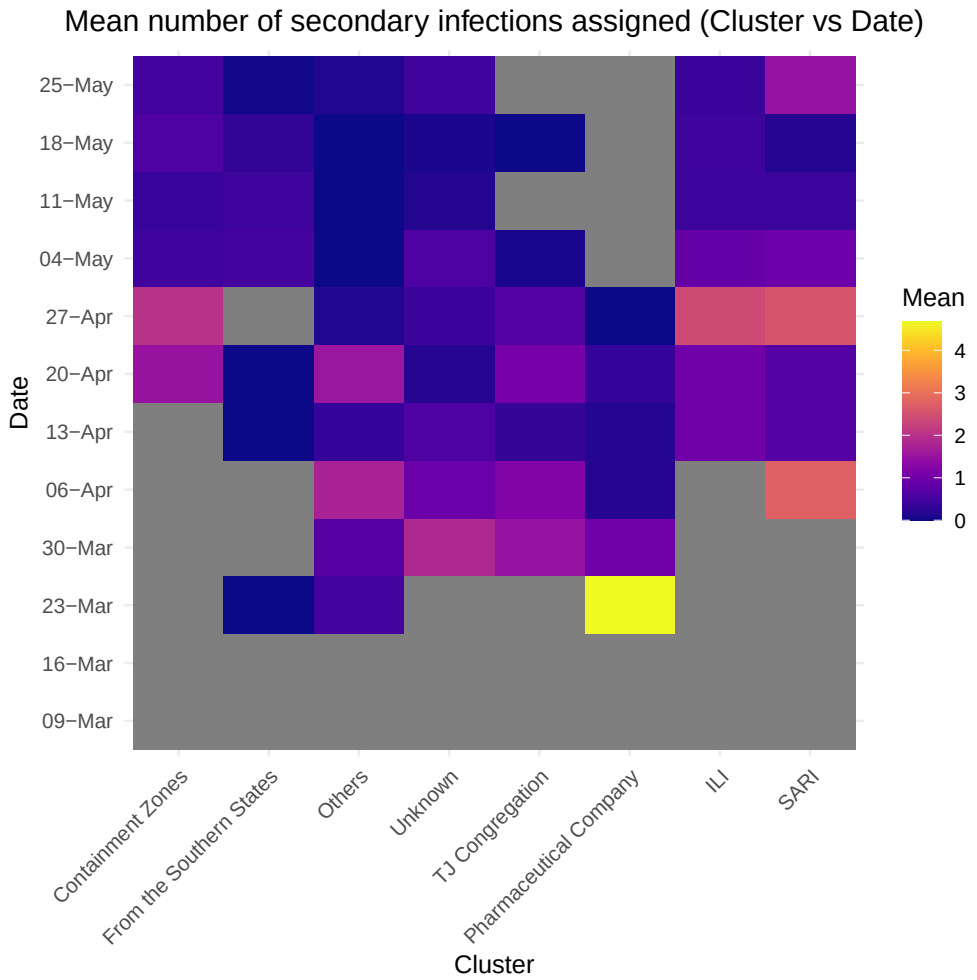


Figure 10: (Cluster across Date) In the above plot, each tile represents a cluster and a week starting at the date given on the left. The color of each tile represents the mean number of infections caused by the patients infected during that week belonging to the cluster. A grey tile means there are no individuals in that cluster who were infected in that particular week.

If we consider 7^{th} April and Descendants till 21^{st} April then there were 290 patients who tested positive and 219 out of them did not pass the infection to anyone else. There was one person who had been assigned 24 secondary infections and the mean number of secondary infections was at 0.6793 with a variance of 4.482 In contrast if we consider the period 7^{th} April to 3^{rd} May and

Descendants till 17 May then there were 615 patients who tested positive and 491 out of them did not pass the infection to anyone else. There was one person who had been assigned 45 secondary infections and the mean number of secondary infections was at 0.7512 with a variance of 9.946.

Age effect in clusters: It is anecdotally believed that socially active patients of median age are the ones who will have higher mean number of secondary infections. We observe that this is in general is not true for some specific clusters. For both SARI and ILI the age group $70-90$ have higher means. This could be because of care takers and close family contracting the infection before the patient tested positive. The TJ Congregation and the Pharmaceutical company clusters both have higher R_{eff} among all clusters. The TJ Congregation has a higher mean across all groups from $10-80$ and the Pharmaceutical company as well has similar features. The meeting attendees as well as the company employees and their families formed bulk of the patients in the respective clusters. Further during that time period media reports suggest that all contacts were being tested and this policy has changed over time.

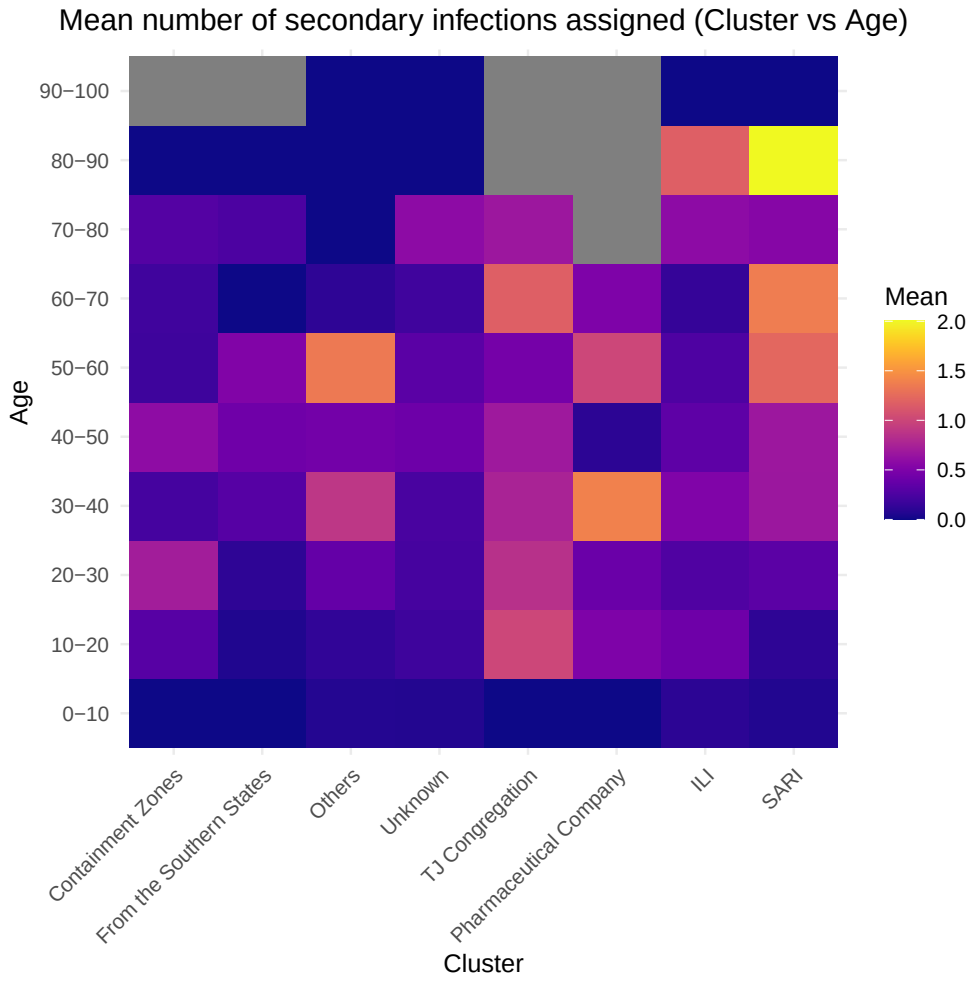


Figure 11: (Ages across clusters) Each tile in the above plot represents a cluster and an age bracket. The color represents the mean number of offsprings for those patients falling into the respective age bracket and cluster. The grey tiles represent the lack of patients falling in that particular demographic.

4 Data

We have sourced all our data from the daily media bulletins of Government of Karnataka. We have converted them from their pdf format and made them publicly available at the Data Repository on our website.

A Model

Let the random variable ν represent the number of infections caused by a particular infected individual, called the individual infectiousness. ν follows a probability distribution whose mean we will designate as R_{eff}. We will assume that

$$\nu \sim \text{Gamma}(k, \frac{k}{R_{\text{eff}}})$$

for some $k > 0$ and

$$Z \sim \text{Poisson}(\nu) \tag{A.1}$$

allowing Z represents the number of secondary infections caused by each infected individual. A standard calculation shows that for $z = 0, 1, 2, 3, \ldots$

$$P(Z = z) = \frac{\Gamma(k+z)}{z!\,\Gamma(k)} \left(\frac{k}{k + R_{\text{eff}}}\right)^k \left(\frac{R_{\text{eff}}}{k + R_{\text{eff}}}\right)^z. \tag{A.2}$$

Thus one interprets Z as having Negative Binomial distribution with mean R_{eff} and Dispersion k[1]. It can also be seen that Z has variance $R_{\text{eff}}(1 + \frac{R_{\text{eff}}}{k})$. Thus smaller the value k, it shall indicate larger variance. Depending on the heterogeneity different models can also be chosen. If one assumed $\nu = R_{\text{eff}}$, then we are assuming a homogeneous population where each individual has the same infectiousness. This will imply $Z \sim \text{Poisson}(R_{\text{eff}})$ $(k = \infty)$ and if we set $(k = 1)$ then $\nu \sim \text{Exponential}(R_{\text{eff}})$, which arises from mean field models assuming uniform infection and recovery rates, will imply $Z \sim \text{Geometric}(R_{\text{eff}})$

B Maximum Likelihood Estimate

Given Data $\mathbf{y} := \{y_i\}_{i=0}^n$, the log-likelihood (modulo constant terms) is

$$L(R_{\text{eff}}, k \mid \mathbf{y}) = \sum_{i=1}^{n} [y_i \log(R_{\text{eff}}) - (y_i - k) \log(1 + \frac{R_{\text{eff}}}{k}) + \sum_{j=0}^{y_i - 1} \log(1 + \frac{j}{k})]$$

We follow [2] to estimate $c = \frac{1}{k}$. First we rewrite the (conventionally accepted) log-likelihood as a function of R_{eff} and $c = \frac{1}{k}$.

$$L(R_{\text{eff}}, c \mid \mathbf{y}) = \sum_{i=1}^{n} [y_i \log(R_{\text{eff}}) - (y_i - \frac{1}{c}) \log(1 + R_{\text{eff}} c) + \sum_{j=0}^{y_i - 1} \log(1 + cj)]$$

[1] In the Epidemiological literature k is referred to as Dispersion and $k > 0$ is assumed, while in the Statistics literature $\frac{1}{k}$ is referred to as Dispersion given the connection with the Gamma distribution and is allowed to take negative values up to $-\frac{1}{R_{\text{eff}}}$

It is then standard (See [7]) that the Maximum Likelihood Estimator for R_{eff} is the sample mean, i.e.

$$R_{\text{eff}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and Maximum Likelihood Estimator for c is a solution to

$$\sum_{i=1}^{n}[\frac{1}{c^2}\log(1+cR_{\text{eff}}) - \frac{y_1 - R_{\text{eff}}}{c(1+cR_{\text{eff}})} - \sum_{j=0}^{y_i-1}\frac{1}{c(1+cj)}] = 0 \tag{B.1}$$

Using (B.1) it is not possible to solve for c explicitly. A numerical approximation scheme is used to obtain an approximate value of c. We use the `uniroot` function in `R`.

C χ^2-goodness of fit test

Given Data $\mathbf{y} := \{y_i\}_{i=0}^{n}$. Let $\hat{R}_0$ and dispersion $\hat{k}$ be Maximum likelihood estimators. To see if Negative Binomial with mean $\hat{R}_0$ and dispersion $\hat{k}$ is a good fit for the data $\mathbf{y}$ we shall perform the χ^2-goodness of fit test. We will consider the range to $\{0, 1, \ldots, B\}$ with $B = \min\{n+1, 20\}$.

$$p_j = \begin{cases} P(Z = j) & \text{for } 0 \le j \le B-1, \\ \\ P(Z \ge B) & \text{for } j = B. \end{cases}$$

Let $y_1, y_2, \ldots, y_n$ be the offspring data from a given cluster and let

$$Z_j = \begin{cases} \#\{1 \le k \le n : y_k = j\} & \text{for } 0 \le j \le B-1, \\ \\ \#\{1 \le k \le n : y_k \ge B\} & \text{for } j = B. \end{cases}$$

Then consider the statistic

$$\mathbf{X}^2 := \sum_{j=0}^{B}\frac{(Z_j - np_j)^2}{np_j} \equiv \sum_{j=0}^{B}\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

As we have estimated two parameters, it is known that $\mathbf{X}^2$- has χ^2_{B-2} degrees of freedom, asymptotically as $n \to \infty$. One way to test if Z is the correct fit for the cluster is to compute the

$$p\text{-value} := \mathbb{P}(\chi^2_{B-2} \ge X^2).$$

There is strong evidence against the possibility that data arose from that model if p-value is very small.

D Confidence Intervals

To compute the confidence interval for the negative binomial dispersion parameter k, we compute it for its reciprocal c and then invert it. We noted earlier that the maximum likelihood estimate

for c had to be solved numerically and it is known that the asymptotic sampling variance is given by a series expansion (See [7]). Let $\hat{c}$ and $\hat{R}_0$ be the M.L.E. obtained. Then let

$$b = \frac{\hat{c}\hat{R}_0}{1+\hat{c}\hat{R}_0} \text{ and } d_i = \prod_{j=0}^{i}(1+j\hat{c}).$$

Then the variance of $\hat{c}$ is given by

$$\sigma^2(\hat{c}) = \left(\frac{n}{\hat{c}^4}\sum_{i=1}^{\infty}\frac{i!(\hat{c}b)^{i+1}}{(i+1)d_i}\right)^{-1}. \tag{D.1}$$

The 95% confidence interval for c is then given by

$$(\hat{c} - z_{0.95}\sigma^2(\hat{c}), \hat{c} + z_{0.95}\sigma^2(\hat{c})),$$

with $z_{0.95}$ being the 95^{th} percentile of the standard normal distribution. The 95% confidence interval for k is then given by

$$\left(\frac{1}{\hat{c} + z_{0.95}\sigma^2(\hat{c})}, \frac{1}{\hat{c} - z_{0.95}\sigma^2(\hat{c})}\right).$$

Note that the above interval will not be symmetric around k due to the inversion. For the computation of Variance in (D.1) we use a tolerance of 10^{-10}.

References

[1] Lloyd-Smith J. Maximum likelihood estimation of the negative binomial dispersion parameter for highly over dispersed data, with applications to infectious diseases. *PloS one*, **12** (2):e180, (2007)

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

[3] Riley S., Fraser C., Donnelly C.A., Ghani A.C., Abu-Raddad L.J., Hedley A.J., Leung G.M., Ho L.M., Lam T.H., Thach T.Q., Chau P. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* **300** (5627), 1961-1966, (2003).

[4] Lipsitch M., Cohen T., Cooper B., Robins J.M., Ma S., James L., Gopalakrishna G., Chew S.K., Tan C.C., Samore M.H., Fisman D. Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300** (5627), 1966-1970. (2003).

[5] Ferretti L., Wymant C., Kendall M., Zhao L., Nurtay A., Abeler-Dörner L., Parker M., Bonsall D., Fraser C. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368** 6491 (2020).

[6] Hellewell J., Abbott S., Gimma A., Bosse N.I., Jarvis C.I., Russell T.W., Munday J.D., Kucharski A.J., Edmunds W.J., Sun F., Flasche S. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health.* (2020).

[7] Saha, Krishna K. and Paul, Sudhir R. Bias-corrected maximum likelihood estimator of the intraclass correlation parameter for binary data, *Stat. Med.*, **24**, 22, 3497-3512, (2005), https://doi.org/10.1002/sim.2197,

[8] Siva Athreya, Nitya Gadhiwala, and Abhiti Mishra, COVID−19 India-Timeline an understanding across States and Union Territories., *Ongoing Study at* http://www.isibang.ac.in/~athreya/incovid19, (2020).

[9] Novel Coronavirus (COVID-19) Media Bulletin, Government of Karnataka, Department of Health and Family Welfare, Bengaluru.https://covid19.karnataka.gov.in/govt_bulletin/en