University of Oxford

Incorporating Household Structure into Branching Processes for Epidemic Propagation: A Novel Multilayered Stochastic Approach with SIR Transmission and Cluster Isolation

by

Anastasia Malakhova

Keble College

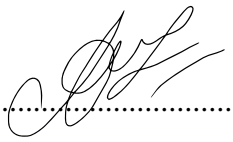
A dissertation submitted in partial fulfilment of the degree of Master of Science in Statistical Science.

Department of Statistics, 24–29 St Giles,
Oxford, OX1 3LB

September 2023

This is my own work (except where otherwise indicated)

Candidate: Anastasia Malakhova

Signed: 

Date:10.09.2023

Abstract

This paper presents a new epidemiological model that is based on a branching process with an addition of household structure to better capture non-random patterns of contact between individuals. The model contains six adjustable parameters - local infection rate within households, global infection rate between households, recovery rate, rate of detection, traceable probability, and distribution of household sizes. We utilise Crump-Mode-Jagers branching processes to analyse the model behaviour in the limit, and rely on simulations to predict an outcome of an epidemic. This enables us to estimate the levels of contact tracing and isolation that are required to put an epidemic under control given initial characteristics of a disease and demographics at the place of an outbreak.

Acknowledgements

First and foremost, I would like to thank my co-supervisors, Prof. Julien Berestycki and Dr. Félix Foutel-Rodier, for the time they invested in offering invaluable suggestions for the direction of my dissertation. I am very appreciative that they met with me multiple times throughout the summer despite a large number of commitments they already had. Additionally, I am thankful for their detailed and constructive comments on my first draft.

I would also like to thank my partner, Grant Sorensen, and my parents, Victor Malakhov and Julia Skornyakova, for providing emotional support and love throughout my time in Oxford. I am forever grateful for that.

Contents

List of Figures

List of Tables

1 Introduction

The study of the spread of epidemics has been a large part of the public health research, especially over the last few years. This is fuelled by the fact that pandemics are now seen as one of the greatest risks faced by humanity. In 2023, World Economic Forum (WEF) has outlined “continued waves of COVID-19” and “Structural failures in health systems” as being among the biggest risks facing the world [7]. Discussions about health crises and potential future pandemics have been part of the agenda on the largest conferences, including United Nations General Assembly (UNGA) and G20 Summits, for many years, and it continues to gain attention. The case of COVID-19, in addition, led to explosion of data available for testing epidemiology models, which led to a rapid increase in the number of papers published on the subject. Even though most models proposed in the papers are oversimplifications of reality, they nevertheless provide insights that can be used for disease control and prevention. They also form a foundation for more complex models leading to an advancement of research in the field. Such insights are crucial when forming prevention strategies, such as the level of isolation and testing in a region, which, in turn, determines public health outcomes.

There are two main ways to model a spread of a disease - through a deterministic or through a stochastic model. Deterministic models have been explored over a number of decades, with "A contribution to the mathematical theory of epidemics" by Kermack, W. O., and McKendrick, A. G. (1927) being one of the most influential papers on the subject [12]. In recent years, however, the availability of increased computing power allowed for an easier application and exploration of stochastic models. Such models tend to be preferred for the analysis of the initial stages of an outbreak, as random events become an important factor when the number of infected individuals is low. They are also useful when there is uncertainty about one or more of the underlying parameters, as they make it possible to incorporate such uncertainty in the analysis. In this work, we focus on the initial stages of an epidemic, and hence concentrate on stochastic models that allow for more realistic representations.

Within each approach, there are further model sub-divisions which depend on the main aim of the analysis and characteristics of the population and disease. As an example, one can use Susceptible-Infected-Recovered or Susceptible-Exposed-Infected-Recovered models which are examples of compartmental models. Both can be used in either stochastic or deterministic setting and both are popular due to their simplicity. Two further prominent approaches involve branching processes and network models. The use of branching processes, such as in [17], is particularly useful when considering the early stages of a spread of a disease and is usually used in stochastic setting. It typically involves each infected individual independently infecting other individuals in an exponential waiting time. On the other hand, network models, such as in [19], allow for a better representation of societal structure as they can imitate non-random patterns of contact between individuals via hyper-edges. This, in turn, makes it easier to incorporate heterogeneity within population. Network models can either be deterministic or stochastic.

The model proposed in this work involves stochastic setting and is inspired by the three papers discussed below. It is based on a branching process, but it additionally incorporates elements of a network model, as individuals are grouped in households. An advantage of

using such model is that it combines an ease and tractability of simulating a branching process, while capturing the effect of household-based societal structure on the spread of a disease. One other advantage of the proposed model is its versatility. It includes six parameters that can be adjusted based on a disease in question, social structure of a population, and the strength of contact tracing and isolation. Overall, the purpose of the model is to be able to set four parameters given a specific disease and demographics, while studying the strength of tracing and isolation required for disease extinction.

In the sections below, we first go through the preliminaries that are required to better capture the dynamics of the discussed models. We then introduce the three works that have inspired the model proposed in this work. Then, the proposed model is outlined and its behaviour described. Further, we analyse what levels and combinations of model parameters would lead to phase transitions (survival vs extinction), as well as provide visual demonstrations of the findings. As part of the analysis, some of our results are compared to the results obtained in the papers that acted as key references in our model development. Towards the end of this work, we discuss the case of COVID-19 in London and present the estimated levels of tracing and isolation that were required in order to hault an outbreak within the timeframe of one year. Finally, we mention the current model limitations and potential improvements.

2 Preliminaries

2.1 Galton-Watson process

A Galton-Watson process is a simple branching stochastic process that was first introduced by F. Galton and H.W. Watson in the study of extinction of family names [25]. The definition of such process, as per [5], is as follows:

Definition 1 (Galton-Watson Process)**.** A Galton-Watson branching process is defined by the probability distribution, $\{p_k\}_{k\in\mathbb{N}}$, of the number of offspring of each node. The process starts with a single node, called the root or the ancestor, and denoted V_0, at generation 0. Then, denote $\eta_{n,i}$ to be the number of offspring of node i in generation n. Then, if X_n denotes the number of nodes in generation n, then

$$X_{n+1} = \sum_{i=1}^{X_n} \eta_{n,i}, \tag{1}$$

where $\{\eta_{n,i}\}_{n,i\in\mathbb{N}}$ are independent and identically distributed with the distribution $\{p_k\}_{k\in\mathbb{N}}$, i.e. for all $k \geq 0$, $\mathbf{P}(\eta = k) = p_k$.

Due to the assumption of independence between nodes and generations, it is easy to see that Galton-Watson process can be easily generalised to other applications, including analysis of spread of a disease. Even though such process would involve many unrealistic assumptions, such as infection rate being the same for all individuals (represented as nodes) across all generations, it provides a good starting point for building a model. In this section, we will quote an important result obtained from analysing such process, and then move on to a more general process that allows for more flexibility and less assumptions.

2.1.1 Galton-Watson tree

A Galton-Watson process can be visualised through the use of a Galton-Watson tree. Such tree is defined as '*an oriented tree spanning the descendants of the ancestor and rooted at the ancestor*', as per page 9 of [5]. Only a single edge between a parent and its child is allowed, and every node has only one parent. The graphical representation of such tree is demonstrated in Figure 1.

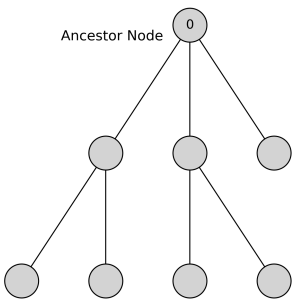


Figure 1: Galton-Watson tree with Poisson offspring distribution with mean equal to 1.5

We can analyse such trees from the perspective of probability of extinction, where extinction refers to an eventual termination of an entire population. Since the process is discrete, such event occurs whenever a single generation produces zero offspring in total, leading to no other nodes that can produce further offspring.

To arrive at the conditions under which we can make probabilistic statements about survival vs extinction of the population, we utilise probability generating functions. Let $\mathcal{T}$ denote a Galton-Watson tree, and $|\mathcal{T}|$ to denote the number of nodes in such tree. Furthermore, let Y denote the number of child nodes that an initial vertex V_0 generates. We then notice that if $\mathcal{T}$ is a tree rooted at V_0, then all of the sub-trees which are rooted at children of V_0, i.e. $\mathcal{T}_1$, ... $\mathcal{T}_Y$, have the same distribution as the initial tree $\mathcal{T}$. Then, as per [5], we define the probability of extinction as below:

$$\begin{aligned} p_{ext} &= \mathbf{P}(|\mathcal{T}| < \infty) \\ &= \sum_{k=0}^{\infty} p_k \mathbf{P}(|\mathcal{T}_1| < \infty, ..., |\mathcal{T}_k| < \infty) \\ &= \sum_{k=0}^{\infty} p_k p_{ext}^k, \end{aligned} \tag{2}$$

We therefore have the following probability generating function of the offspring distribution:

$$\phi_\eta(s) \coloneqq \sum_{k \in \mathbb{N}} p_k s^k, \tag{3}$$

with p_{ext} being equal to the solution of the below equation:

$$x = \phi_\eta(x) \tag{4}$$

It can further be shown that the probability of extinction would correspond to the smallest root of $x = \phi_\eta(x)$ in [0,1]. Then, if we denote μ as the average number of children, one can use the fact that $\phi'_\eta(1) = \mu$ and that ϕ_η is non-decreasing and convex, in order to prove Theorem 1.2 of [5], which we quote below:

Theorem 1 (Survival vs extinction)**.** The extinction probability p_{ext} is the smallest solution of equation [4]. Denoting by $\mu \coloneqq \mathbf{E}(\eta)$ the average number of children per individual, one further has the following regimes:

- *Subcritical regime: If* $\mu < 1$*, then* $p_{ext} = 1$
- *Critical regime: If* $\mu = 1$ *and* $p_1 < 1$*, then* $p_{ext} = 1$
- *Supercritical regime: If* $\mu > 1$*, then* $p_{ext} < 1$

Such theorem is important, as it defines conditions for the only possible outcomes of the process in the long-run. We either have an eventual extinction for sub-critical and critical cases, or a survival of the process. Furthermore, with a more granular analysis of the supercritical case, it is possible to show that if we condition on the survival of the process, this would lead to a geometric growth of the population. The only exception to a geometric growth given survival is the case where each individual gives birth to exactly

one individual with probability 1. In this case, we have survival of the population, with the population total remaining constant. This theorem is useful when considering a spread of a disease, as we can estimate μ empirically. This, in turn, helps us navigate the strength of isolation strategies needed to halt an epidemic. Furthermore, the above result allows us to get insights into the types of long-term outcomes we should expect when considering more complex processes, as we will see below.

2.2 Crump-Mode-Jagers process

Crump-Mode-Jagers (CMJ) process is a continuous time generalisation of a Galton-Watson process. As part of this process, individuals have random lifetimes, during which they produce offspring according to a point process. Such point process may have a varying rate depending on the age of an individual. In epidemiological models, for example, one of the well known point processes, denoted by ξ and defined on $(0, \infty)$, is a Poisson point process, where $\xi(t)$ represents the number of offspring produced by an individual up to time t. We additionally note that we use ξ to denote both, a distribution function and a measure. We then note that such process has atoms as integer-valued weights. Furthermore, the definition of CMJ relies on a pair of random variables, (V, ξ), which can be dependent. For that pair, V_i denotes a real-valued random variable that is seen as a length of life of an individual i.

A formal definition of such process, as per [15], is as follows:

Definition 2 (Crump-Mode-Jagers Process)**.** A CMJ process is a branching process that is defined through a pair (V, ξ), where ξ is defined on $(0, \infty)$. As part of the process, each individual i is given an independent copy (V_i, ξ_i) of (V, ξ). If individual i is born at time B_i, then at time t, where $t \geq B_i$ and $t \leq B_i + V_i$, she gives birth to $\xi_i(\{t - B_i\})$ i.i.d. copies of herself. If ξ is a Poisson process independent of V then the CMJ process is called binary and homogeneous.

As can be seen from the definition above, a CMJ process allows for heterogeneity between individuals, such as individuals having a random length of life, and a possibility of having different offspring distribution for different points in time. Relaxation of such assumptions compared to the Galton-Watson process allows for more realistic models, and hence is applicable for the proposed model in this work, as will be shown below. It is important to mention, however, that the total offspring distribution remains the same for all individuals – such concept would be important for the calculation of the Malthusian parameter. Furthermore, as per [23], all individuals have the same birth and death distributions regardless of time. Such assumption is realistic whenever we consider a spread of a single disease without accounting for mutations that may change the severity and length of illness. In this work, we concentrate on the early stages of epidemics, and therefore it is a reasonable assumption to make.

An important consequence of a process being a CMJ process is an ability to determine its long-term behaviour based on the expected number of children. Given the above notation, the expected number of children at time t can be expressed as $E[\xi_i(\{t-B_i\})]$. Using [10], we then arrive at the following limiting behaviours:

- Supercritical case: $E[\xi([0,\infty))] > 1$
- Critical case: $E[\xi([0,\infty))] = 1$ and $P(\xi([0,\infty)) = 1) < 1$
- Subcritical case: $E[\xi([0,\infty))] < 1$

Through further non-trivial analysis of the supercritical case, it can be shown that supercritical case leads to an exponential growth of the process. Critical and subcritical cases, on the other hand, lead to an almost sure extinction.

2.2.1 Malthusian parameter

One of the most used concepts in epidemiology undermining spread of disease models involve Malthusian parameter, which represents an average growth rate of an epidemic. Let us denote such parameter by α. In order to define it, as per [10], we first consider the random point process ξ_i for an individual i. We note that $\xi_i(0) = 0$, and that ξ_i increases by increments of 1 at the times that individual i gives births. Then, given the same law of reproduction for every individual, we can define $\mu(t) = E[\xi(t)]$. As a result, we define a Malthusian parameter α such that:

$$\int_0^\infty e^{-\alpha t} d\mu(t) = 1 \tag{5}$$

The Malthusian parameter is an important quantity in branching processes, since its sign creates classification into subcritical, critical and supercritical regimes. More precisely:

- if $\alpha > 0$, the number of infected individuals is expected to grow exponentially
- if $\alpha < 0$, the number of infected individuals is expected to drop to 0
- if $\alpha = 0$, the number of infected individuals remains constant

2.3 Other useful definitions

It is also worth defining a range of epidemiology-related terms that would be used throughout the paper.

Basic Reproduction Number (BRN or R_0) - The basic reproduction number implies the number of secondary infections in an otherwise susceptible population caused by a single original infection. [1]

Note that when applied to susceptible population inside one household, it would be called Household Reproduction Number (HRN).

Secondary Attack Rate (SAR) - The number of cases of an infection that occur among contacts within the incubation period following exposure to a primary case in relation to the total number of exposed contacts; the denominator is restricted to susceptible contacts when these can be determined. [21]

Epidemic - A sudden outbreak of infectious disease that spreads rapidly through the population, affecting a large proportion of people. [16]

Endemic - In epidemiology, referring to a disease that occurs continually in a given geographical area. [13]

3 Literature Inspiring the Proposed Model

3.1 'A Growth-Fragmentation-Isolation process on random recursive trees and contact tracing' by Vincent Bansaye et. al (2022)

The model proposed in this work is inspired by the three papers, each of them involving a stochastic setting. The first paper, [24], considers a stochastic process on recursive trees. The process starts at an active root vertex V_0 at time $t = 0$. The overall process can then be characterised via considering a dynamic tree $G(t) = (V(t), E(t))$, and two functions – $\Psi_t : V_t \to \{0, 1\}$, and $\eta_t : E_t \to \{0, 1\}$. $\Psi_t(v)$ indicates whether a vertex v is active (1) or inactive (0), while $\eta_t(e)$ indicates whether an edge e can be traced (1), or not (0). Each active infected individual is therefore represented as a node which gives birth to other nodes in an exponential waiting time with parameter γ. When such infection occurs, a child vertex is labelled with $v \in [0, \infty)$, i.e. the time it was infected. Furthermore, once infection occurs, an edge that is created between a parent and a child node is initially labelled as traceable. However, those edges can become untraceable later on in the process, which happens in an exponential waiting time with parameter β. This leads to a fragmentation of a tree and appearance of clusters. We note that in this case a cluster would correspond to a separate connected component in a graph where we only keep traceable edges and remove all untraceable edges. Finally, clusters are isolated at a rate proportional to their size.

The process described above is demonstrated in the graphs below, where we assume for simplicity that we already have six distinct nodes which were born at times $t = 1, 2, 3, 4, 5$ and a root node at time $t = 0$. We then add another node at time $t = 6$ in Figure 2, indicating Growth.

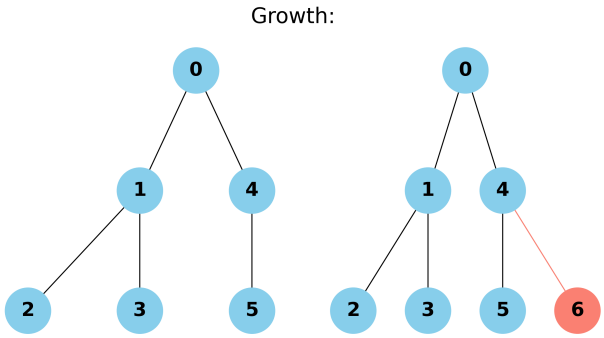


Figure 2: Node 4 giving birth to node 6 with an initially traceable edge

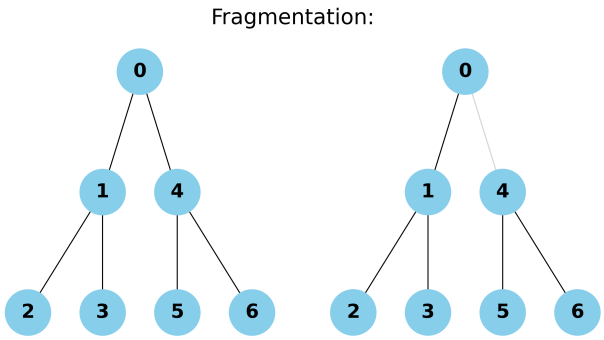


Figure 3: Fragmentation and appearance of two clusters: 0,1,2,3 and 4,5,6

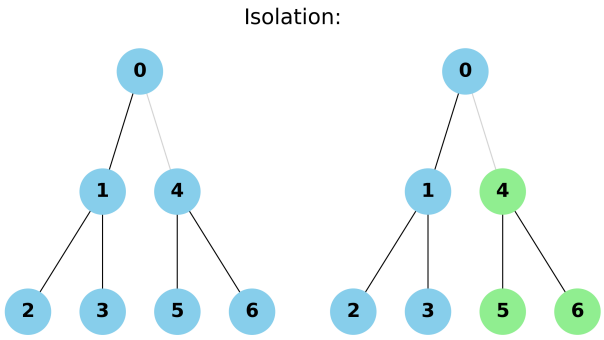


Figure 4: Isolation of cluster 4,5,6

The dynamic described above leads to the growth-fragmentation-isolation process.

In order to analyse its behaviour and explore phase transitions, the authors analyse behaviour of clusters that appear due to fragmentation. In order to work with such clusters, the Ulam-Harris-Neveu labelling is utilised which is described below, as per [24]:

- An initial cluster is labelled as $\emptyset$ by convention
- Each further cluster has a unique label u that is part of a set U, where U is defined as follows:

$$U = \bigcup_{n \geq 0} \{1, 2\}^n \tag{6}$$

- Once fragmentation occurs, a cluster is split into two child nodes, $u1$ and $u2$. The first child, $u1$, would be the one containing the root of a cluster, where the root is defined as a vertex with a minimum infection time label.

When considering the whole system, the authors note that it is easier to analyse the process via considering a cluster process $(\mathcal{X}, \mathcal{Y})_{t \geq 0}$, where:

$$\begin{aligned} \mathcal{X} &= \{C \mid C \text{ is a cluster in } G_t;\ \forall v \in C,\ \Psi_t(v) = 1\} \\ \mathcal{Y} &= \{C \mid C \text{ is a cluster in } G_t;\ \forall v \in C,\ \Psi_t(v) = 0\} \end{aligned} \tag{7}$$

If a cluster is active, i.e. is in set $\mathcal{X}$, then each node within such cluster continuously attaches new nodes in an exponential waiting time. If, on the other hand, a cluster is inactive, i.e. in $\mathcal{Y}$, then it is considered isolated and will not produce any more offspring. As a result, each cluster that is created through the process of fragmentation can be described via the random recursive tree structure.

Such description of the model through active and inactive clusters would be useful for us when analysing the model proposed in this work.

3.2 'A model for an epidemic with contact tracing and cluster isolation, and a detection paradox' by Jean Bertoin (2022)

Similar to [24], the second paper focuses on a scenario where individuals are either infected or isolated, i.e. it doesn't consider recovery or death. Furthermore, each newly infected individual can either be traceable to its parent, or untraceable. This additional feature allows the author to analyse model's dynamics from the perspective of clusters of infected nodes rather than individuals. It is worth highlighting that traceability of an edge is decided at a point when such edge is created, and it doesn't change its status at any further time. As a result, it is different from the fragmentation dynamics used in [24].

The model starts with a single infected individual, and follows the below dynamics:

- Each infected individual infects other individuals with rate γ.
- At an instance of infection, an edge is created as traceable with probability $p \in (0, 1)$, or untraceable with probability $1 - p$.
- Each individual gets detected at rate δ. Once an individual gets detected, the whole cluster that individual belongs to is isolated, where cluster refers to a collection of nodes with only traceable edges between them.

An important point is the definition of generations. Since analysis is conducted on a cluster-level, the generation number is defined as the number of untraceable edges between that cluster and the ancestor.

The below diagram demonstrates the model more clearly:

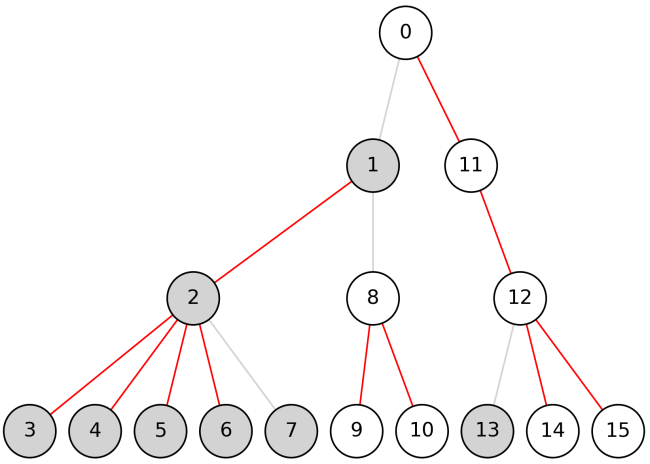


Figure 5: Illustration of Jean Bertoin's model

In Figure 5, red lines represent traceable edges which form clusters. The ancestor is at the top, denoted by "0". There are a total of five clusters: $\{0, 11, 12, 14, 15\}$, $\{13\}$, $\{8, 9, 10\}$, $\{7\}$ and $\{1, 2, 3, 4, 5, 6\}$. Furthermore, two clusters, $\{0, 11, 12, 14, 15\}$ and $\{8, 9, 10\}$, are isolated, while the rest remain active. Cluster $\{1, 2, 3, 4, 5, 6\}$ is considered to be a first generation cluster, and cluster $\{8, 9, 10\}$ is considered to be a second generation cluster.

The above results in the clusters following a Crump-Mode-Jagers branching process. It starts at an ancestral cluster $\mathbf{C}_\emptyset$, and continuously adds new clusters with the same underlying evolution process $(\mathbf{C}_u)_{u\in U}$. Such process is defined as follows:

$$\mathbf{C}_u = (C_u, \xi_u), \qquad u \in U, \tag{8}$$

where $C_u(t), t \geq 0$ represents the size of a cluster at time t, with $C_u(t) = 0$ after detection and isolation of that cluster. The $\xi_u([0, t])$ represents the number of child clusters that a cluster u has generated since birth up to time t. Such notation is useful for us, as we will also analyse the dynamics of our model through utilising clusters instead of directly working with individuals.

The main analysis and results are given for the super-critical case, i.e. conditioned on the disease's survival. For this scenario, the distribution of the isolated cluster sizes are considered. Since there are only contamination (γ) and isolation (δ) dynamics, Bertoin's paper mentions that a typical cluster at a point of isolation will follow a geometric distribution, i.e.

$$P(size = k) = \frac{\delta}{\delta + p \times \gamma}(1 - \frac{\delta}{\delta + p \times \gamma})^{k-1}, \tag{9}$$

where $P(size = k)$ represents probability that a cluster will have size k at isolation.

The author notes, however, that the distribution of isolated cluster sizes differs from the above geometric distribution when considered in the limit, as $t \to \infty$. More precisely, if α represents Malthusian parameter that depends on the underlying parameters γ, p, δ, then we get:

$$P(size = k) = c_i(1 - \frac{\delta}{\delta + p \times \gamma})^{k-1} B(\frac{\alpha}{\delta + p \times \gamma}, k + 1), \tag{10}$$

where c_i is the normalisation factor, and B is the beta function.

The author highlights that obtaining two different distributions for the sizes of clusters at a point of their isolation, and in the limit, is surprising, and refers to it as a 'detection paradox'.

Let us consider the calculations in more detail. In Bertoin's paper, in order to derive the measure for the active clusters of size k, the author utilises the known properties of Malthusian parameter, which was introduced in section 2, together with Laplace transform. More concretely, define $\mu(t)$ to be the intensity measure of ξ:

$$\mu(t) := E[\xi([0,t])], \quad \text{where} \quad t \geq 0 \tag{11}$$

Then, the author expresses how quickly untraceable contaminations occur in a typical cluster through the following transform:

$$\mathcal{L}(x) = \int_0^\infty e^{-xt} d\mu(t), \quad x \geq 0 \tag{12}$$

Given that the author specifically considers the case where probability of survival is non-zero, only cases where $(1-p)\gamma > \delta$ are considered.

The Malthusian parameter can then be expressed via a unique solution $\alpha = \alpha(\gamma, p, \delta)$ of the equation $\mathcal{L}(x) = 1$. Then, given Corollary 3.4 of the paper, which states that:

For every $t \geq 0$, there is the identity

$$\mu(t) = (1-p)\frac{\gamma}{\delta}(1 - \frac{1}{1+\delta(e^{\rho t}-1)/\rho}) \tag{13}$$

where $\mu(t) := E[\xi([0,t])]$, with $t \geq 0$, $\rho = \delta + p \times \gamma$ and ξ being a point process described earlier in the section, the author is able to re-write the above transform as

$$(1-p)\gamma \int_0^\infty \frac{e^{(-\alpha)t}}{(1+\delta(e^{\rho t}-1)/\rho)^2} dt = 1 \tag{14}$$

Furthermore, J. Bertoin introduces measures for active and isolated clusters, m^a and m^i that are written as

$$< m^a, f >= \int_0^\infty e^{\alpha t} \times E(f(C(t)), t < \zeta) dt, \tag{15}$$

and

$$< m^i, f >= \int_0^\infty e^{\alpha t} \times E(f(C(\zeta-)), \zeta \leq t) dt, \tag{16}$$

Such measures represent an exponentially weighted expected value of a function f, applied to the size of a typical cluster at a time prior to its isolation for active clusters, and post its isolation for inactive clusters. In this notation, f represents a generic non-negative function.

Finally, by utilising the definition of the Beta function, the author arrives at the following expression for the measures of active and inactive clusters via the change of variables:

$$\begin{aligned} m^a(k) &= (1-\frac{\delta}{\rho})^{k-1}\int_0^\infty e^{-\alpha t}(1-e^{-\rho t})^{k-1}e^{-\rho t}dt \\ &= \frac{1}{\rho}(1-\frac{\delta}{\rho})^{k-1}B(1+\frac{\alpha}{\rho},k), \end{aligned} \tag{17}$$

and

$$\begin{aligned} m^i(k) &= \frac{\delta}{\rho}(1-\frac{\delta}{\rho})^{k-1}\int_0^\infty e^{-\alpha t}(1-e^{-\rho t})^{k}dt \\ &= \frac{\delta}{\rho^2}(1-\frac{\delta}{\rho})^{k-1}B(\frac{\alpha}{\rho},k+1), \end{aligned} \tag{18}$$

which has the form presented earlier for the limiting distribution of the sizes of clusters.

3.3 'Household Epidemic Models and McKean-Vlasov Poisson Driven Stochastic Differential Equations' by Raphaël Forien and Étienne Pardoux (2022)

R. Forien and E. Pardoux's epidemiological model is quite different from the two models described previously. First of all, their model is analysed with an assumption of a finite population size. Additionally, it introduces a two-level mixing model, where individuals are placed in households. Whenever an infection occurs between such households, an individual that gets infected is chosen uniformly from a population, which is the same as selecting a household from a size-biased distribution, and then uniformly selecting an individual inside that household. In addition, for within households infections, a Susceptible-Infected-Susceptible (SIS) process is followed, where an individual doesn't retain any immunity. We straight away note that this approach is aimed at analysing endemic scenarios, and is therefore concerned with different quantities of interest, such as proportion of infected individuals in the limit.

The model is defined on $\mathcal{N}$ households, and infection rates differ for within (λ_L), and between (λ_G) household transmissions. Given that it involves an SIS model, each individual recovers at rate γ, which is i.i.d. exponential. This results in a Markovian dynamic, where each next event is only dependent on the current state of the system. Unlike previous papers, the model doesn't incorporate tracing or isolation, so the only opposing force to the growth of infectious individuals is recovery.

The main question that the paper is trying to answer refers to the distribution of the asymptotic number of infected individuals in a typical household, as $t \to \infty$ and $\mathcal{N} \to \infty$. It uses the fact that the interactions between each household is of the mean-field type, which means that the infection rate can be expressed using the average number of infected individuals in all households.

In order to describe the number of infectious individuals per household, the authors use stochastic differential equations of the following form, as in the Definition 2.1 of [6]:

$$
\begin{aligned}
X_i^{\mathcal{N}}(t) &= X_i(0) - P_{rec,i}\Big(\gamma \int_0^t X_i^{\mathcal{N}}(s)ds\Big) + \\
&P_{\text{inf},i}\Big(\int_0^t \big(1 - \frac{X_i^{\mathcal{N}}(s)}{\nu_i}\big)\big[\lambda_L \frac{\nu_i}{\nu_i - 1} X_i^{\mathcal{N}}(s) + \lambda_G \frac{\nu_i}{\overline{\nu}^{\mathcal{N}}} \frac{1}{\mathcal{N}} \sum_{j=1}^{\mathcal{N}} X_j^{\mathcal{N}}(s)\big] ds\Big),
\end{aligned}
\tag{19}
$$

where $X_i^{\mathcal{N}}$ is the number of infected individuals in household i, and there are $\mathcal{N}$ households at time t in the model; $(P_{rec,i}(t), t \geq 0, i \geq 0)$ and $(P_{inf,i}(t), t \geq 0, i \geq 0)$ are Poisson processes representing the number of recovered and infected individuals, respectively, in household i.

The main analysis of the paper is quite different from the questions we will try to answer in this work. Nevertheless, it is useful for us, as it provides ideas on how one can represent societal structure when modelling the spread of a disease.

4 Model Description

4.1 Model Definition

The model proposed and analysed in this work attempts to add an element of societal structure into the branching process in order to make it more realistic. It does so through creating a multiple level structure, where individuals are grouped into households, and households are the ones representing the nodes in the branching process. Furthermore, individuals infect each other within households, as well as between households, resulting in growth of a tree. The between-households infections are applied to an infinite pool of susceptible individuals, and we additionally note that we analyse the model from the perspective of an early stage of an epidemic.

We now describe the formal definition and dynamics of the proposed model.

Definition 3 (Proposed model)**.** Define a dynamic tree $G_t = (V_t, E_t)$, where V_t is a set of vertices representing households, and E_t represents edges between those households, indicating a relationship between the parent node (infecting household) and child nodes (new households getting infected). Furthermore, let each household consist of a number of sub-vertices, i.e. individuals. If we assume that each individual is equally likely to get infected, then the size of the household of that individual is size-biased, with $P(h_k = i) = \pi_i^* = \frac{i*\pi}{\sum i*\pi}$.

Define a stochastic growth-isolation process with fixed-status edges on a dynamic tree. Start with a household V_0 at time $t = 0$, and one infected individual in such household. Then, follow the below process:

Growth: Each household independently attaches new households with one infected individual. It does so at rate $k \times \beta$, where k is the number of infectious individuals in the household, and β is the global infection rate. Additionally, each infected individual continuously infects other individuals in the same household at a local infection rate λ, and each infected individual recovers at a rate γ, resulting in SIR process.

Fixed-status edges: Each time a new household gets infected, an edge between a parent and child nodes is created as either traceable, with probability p, or untraceable, with probability $1 - p$. When an edge is created as untraceable, it leads to an appearance of a new cluster, where a cluster of nodes is defined as a set of households all of which have traceable edges between them, i.e. there exists a traceable path between any two nodes in the same cluster.

Isolation: A confirmation process is such that each infectious individual (the sub-node inside households) gets confirmed in an exponential waiting time with parameter θ. Once an individual is confirmed, the whole cluster of households this sub-vertex belongs to is isolated. Isolation implies that all households in that cluster can no longer produce offspring.

We therefore arrive at the following parameters that can be varied in order to analyse properties of the above process:

- π - distribution of household sizes (with π_* being a size-biased distribution)
- λ - local infection rate
- γ - local recovery rate
- θ - confirmation rate of infected individuals
- β - global infection rate
- p - probability of generating a traceable edge

We note that π, λ, γ and β, relate to the initial stages of an epidemic, and can be estimated based on the specifics of the disease and population demographics. The other two parameters, notably θ and p, relate to the strength of policies that can be implemented in the region in an attempt to stop a disease from spreading. The confirmation rate would correspond to the strength of testing, while probability of generating a traceable edge – to the strength of contact tracing. The first four parameters can therefore be estimated based on transmissibility data and demographics from the epicenter location. Based on those, one can then find the strength of testing and isolation required to lead a disease to extinction in a reasonable time.

We additionally highlight that we incorporate fixed-status edges as opposed to fragmentation, as we believe human behaviour to be a larger influential factor in contact traceability. More concretely, there is more variability in people deciding to either follow rules and guidelines (and hence report their interactions), or not following them, compared to loosing a data point that was already obtained.

Given the above specification, our model is a continuous-time, discrete state stochastic process, as we generate time between events as a continuous random variable, and the number of individuals, number of households, and relationships between households are represented as non-negative integers. Furthermore, there are three main levels that the model can be analysed at. The lowest level of the process consists of individual subvertices in a household which can either be infected, recovered, or susceptible. The second level involves households that follow a branching process. Finally, the third level consists of clusters of households. We will show that through utilising the third level, our proposed model can be presented through the lens of Crump-Mode-Jagers process, similar to the analysis used in [2]. Presenting our process in this way will make it easier for us to analyse the model's behaviour in the limit, as $t \to \infty$, as we will be able to use the known properties of CMJ processes. Let us now consider the three levels in more detail to gain an intuition into the dynamics of the process.

4.2 The three levels of the proposed model

We start by highlighting that each individual is infecting other individuals within the same household independently of others. This notion further translates into independence between households, where each household's infection rate depends on the number of in-

fected individuals within that household and global rate β, however, is independent of any other household. Finally, such independence also holds for clusters of households.

Additionally, we will use Ulam-Harris notation, similar to the notation used in [2], for individual households in the 4.2.2 section, and for clusters in the 4.2.3 section. More precisely, we use the following notation for identification of households:

- An initial household is labelled as $\emptyset$ by convention
- Each new household has a unique label k that is part of a set K, where K is defined as:

$$K = \bigcup_{n=0}^{\infty} \mathbb{N}^n \tag{20}$$

- A household k generates new households labelled $k1$, $k2$, ... based on the increasing order of birth times of such households. The birth of a new household occurs when household k gives birth to a new household with either traceable or untraceable edge

Then, the following notation is used for the identification of clusters:

- An initial cluster is labelled as $\emptyset$ by convention
- Each new cluster has a unique label u that is part of a set U, where U is defined as:

$$U = \bigcup_{n=0}^{\infty} \mathbb{N}^n \tag{21}$$

- A cluster u generates new clusters labelled $u1$, $u2$, ... based on the increasing order of birth times of such clusters. The birth of a new cluster occurs when cluster u generates a new household with an untraceable edge

4.2.1 First level: Individuals

Let us consider the first level of the model, i.e. individuals inside a single household. We straight away notice that if we put an upper bound on the number of individuals that can be present in a single household, we have a finite-dimensional state space for the number of susceptible (S_k), infected (I_k), and recovered (R_k) individuals. The subscript k in this case would correspond to an identification number of a single household, where $k \in K$. Let us also ignore detection and global infections at this level, which would lead us to the system following a Susceptible-Infected-Recovered dynamic. In this case, the system can be represented via $(S_k(t), I_k(t), R_k(t)) \in \mathbb{N}^3$, with $S_k(0) = h_k - 1, I_k(0) = 1$, and $R_k(0) = 0$ at the time of birth of household k, where h_k is the size of household k. We additionally have a constraint of $S_k(t) + I_k(t) + R_k(t) = h_k \ \forall t \in [0, \infty)$. Then, we can define transition rates as follows:

$$\begin{cases} (S_k, I_k, R_k) \to (S_k - 1, I_k + 1, R_k) & \text{with rate } \lambda I_k(\frac{S_k}{h_k - 1}) \times \mathbb{1}\{h_k > 1\} \\ (S_k, I_k, R_k) \to (S_k, I_k - 1, R_k + 1) & \text{with rate } \gamma I_k \end{cases} \tag{22}$$

An indicator function inside the transition rate for an increase in the number of infected individuals ensures that if we have a household consisting of a single individual, then infection rate is 0 – there is no one else to infect.

4.2.2 Second level: Households

Let us now extend this representation to a scenario where we consider multiple households, i.e. the full model where households represent nodes in a branching process. Since we assume an infinite pool of susceptible individuals, we only focus on infected individuals and inter-dependencies between households. In this case the Markov Chain has infinite state space.

To describe such scenario, let $V(t)$ represent a set of households that are present and active at time t. We note that such set would consist of identification numbers, $k \in K$, as per the Ulam-Harris notation. We then have:

$$\begin{cases} I_k \to I_k + 1, & \text{at rate } \lambda I_k \frac{S_k}{h_k - 1} \times \mathbb{1}\{h_k > 1\} \\ I_k \to I_k - 1, & \text{at rate } \gamma I_k, \end{cases} \tag{23}$$

for $k \in V(t)$. Furthermore, a global infection happens at rate $\beta \sum_{k \in V(t)} I_k$. In case of a global infection, a parent node is chosen in such a way, that the probability of a node being selected as a parent is equal to $\frac{I_k}{\sum_{k \in V(t)} I_k}$. Once infection occurs, a household identification number of a new node is added to set $V(t)$.

We then additionally need to consider the case of detection and isolation. As per the above model description, once detection occurs, the confirmed node is being isolated together with the rest of the cluster. Hence, isolation rate should take the whole cluster into account. In [2], the author notices that a typical size of a cluster at a point of isolation follows a geometric distribution. However, in our case, each subsequent event that applies to a cluster depends on the previous state of this cluster. More precisely, given that confirmation rate per infected individual is θ, and the rate of generating a traceable birth per infected individual is $p \times \beta$, we arrive at the following probability of the next event happening to a cluster being detection:

$$\frac{\theta \sum_{k \in C} I_k}{\theta \sum_{k \in C} I_k + \gamma \sum_{k \in C} I_k + p\beta \sum_{k \in C} I_k + \lambda(\sum_{k \in C}(I_k(\frac{S_k}{h_k - 1})) \times \mathbb{1}\{h_k > 1\}}, \tag{24}$$

where C denotes the set of households that are part of the cluster. What we can therefore observe, is that the probability that the next event is detection depends on the current state of the cluster, and hence we are not able to use the desirable properties of geometric distribution to describe the typical cluster size at a point of isolation. This means that there is no simple expression for the average number of individuals that get isolated once a detection event occurs. Hence, in order to account for isolation, we need to incorporate the information on traceability between each household in the set $V(t)$, which can be presented as another set. Such notation is more complex, however, so we instead focus on analysing the third level of our model - clusters of households.

4.2.3 Third level: Clusters of households

In order to describe the model as a CMJ process, we will utilise the same notation as in [2]. First, we remind the reader that a cluster C generates child clusters whenever any household $k \in C$ gives birth with an untraceable edge. Given independence criterion between households, and, subsequently, clusters of households, one can see that each cluster can then be described by a pair $\mathbf{C}_u = (C_u, \xi_u), \ \ u \in U$, where U is a set of identification numbers for clusters, $C_u(t)$ represents the size of a cluster at time t, and $\xi_u([0,t])$ represents the number of child clusters generated by cluster u since birth and up to time t. Furthermore, we define ζ_u to represent the age at which a cluster becomes inactive, i.e. its length of life. We note that unlike Bertoin's model, a cluster can become inactive in two ways - either through detection and isolation, or through all individuals successfully recovering inside the cluster. As a consequence, $\xi_u([\zeta_u, \infty)) = 0$, since once the cluster becomes inactive, it no longer produces any offspring, traceable or untraceable. Since we are interested in the number of infectious individuals that continue spreading the disease, we further assume that $C(t) = 0$ for $t \geq \zeta_u$, and define $C(\zeta_u -)$ to represent the size of a cluster at an instant before its isolation or recovery of all individuals within it.

Given the above, we can make two observations. First, the distribution of a cluster at the root will correspond to the distribution of all clusters in the generations that follow. Second, we can therefore view such process as a one-type CMJ process.

As a result, given the long-term behaviours introduced in section 2.2 section, we know that the limiting behaviours of the process, as $t \to \infty$, can only include an almost sure extinction or an exponential growth (or constant population size if $P(\xi_u([0,\zeta_u]) = 1), \ \forall u \in U$). Additionally, the average number of child clusters can be seen as a Basic Reproduction Number (BRN), determining the outcome of the process. It is worth noting that the model doesn't have any generational dependency, however, we do have dependency on the age of a cluster. As a result, if one is estimating BRN using simulations, one has to make sure that clusters are considered in the calculation only once they become inactive as part of the process, resulting in a difficulty of obtaining an unbiased estimate. In a later section, we look at a number of ways one can approach such estimation.

5 Simulations for phase transitions analysis

In this section, we explore the relationships between model parameters and how they affect model dynamics and threshold behaviour. The model is first initialised with default parameters which are listed below. The values of the parameters are then altered to further explore the relationships between them and the outcome of simulations.

5.1 Default parameters

In real world, there are many factors that can affect the resulting values of parameters relating to disease transmission. For example, the rates at which individuals infect each other would depend on the disease in question, how often such individuals interact, and location of interaction (e.g. indoors or outside). It is therefore difficult to pinpoint exact values of parameters that can be used as defaults in simulations. Here, we are using some general guidelines and empirical research results in order to arrive at reasonable parameter defaults.

Time units: In this work, we are assuming that one unit of time corresponds to one day.

Local infection rate: Local infection rate refers to the rate at which individuals infect each other within each household. It is therefore reasonable to assume that individuals are going to be in close proximity to each other. Given such conditions, it is plausible that individuals would infect each other every couple of days. Hence, we start with a local infection rate of 0.5.

In case of COVID-19, [20] and [14] suggest that a local infection rate is significantly greater than a global infection rate, and point towards heterogeneity in Secondary Attack Rate estimates across a number of studies. In contract, the results of studies on influenza A (H1N1), such as in [3], point towards lower transmission rates for this disease. Given that our model is a general model that can be applied to a variety of diseases, we consider a relatively wide range for the local infection rate to be between 0.02 and 1.

Global infection rate: Global infection rate would refer to individuals infecting other individuals from different households. Such infections would therefore happen either outside, at work or in public places like restaurants. We take global infection rate to be smaller than the local infection rate, at 0.1. Additionally, we set the range to be between 0.05 and 0.25 to reflect the differences between neighbourhood structures, traditions and time proximity to public holidays.

Local recovery rate: This parameter will primarily depend on the type of a disease and availability and quality of health services provided. It is worth noting that our model uses SIR dynamic, which means that once recovered, an individual doesn't become susceptible once more. This means that we consider cases of either constant immunity or death. Here, we take a local recovery rate to be 0.15, i.e. based on an average length of a common influenza disease [9]. Since average length of a disease will vary substantially based on the disease in question, we consider a range of 0.05 to 1.

Confirmation rate: This rate refers to how quickly we are able to identify that a certain person has a disease, and how quickly we are able to isolate this individual. Such rate

may depend on the overall efficiency of health services in a country, as well as how fast a person is likely to develop and identify symptoms given a particular disease. In this work, we consider the beginning stages of a spread of a disease, which means that confirmation rate is likely to be slow. We set confirmation rate at 0.05. The range of values considered are between 0 and 1.

Distribution of the number of individuals in a household (size biased): We assume that our population consists of households with up to five people per household. We further assume that there are equal number of such households of each size. The distribution in a simulation is size biased, as each individual is five times more likely to infect another household which has five individuals compared to a household with only one individual.

Probability of a traceable edge: Probability of a connection between a newly infected person and an individual who infected this person would depend on the strength of tracing in the location where infection occurred. We take probability of a traceable edge to be 0.06, based on the estimates in [26] and [11]. For COVID-19, for example, the strength of tracing varied from no tracing at all, as in Sweden [4], to strongly enforced tracing, as in China. We therefore consider the whole range of probabilities from 0 to 1 for this parameter.

Since branching processes are typically used when analysing the initial stages of an epidemic, we focus on the spread of a disease prior to the population reaching 100,000 individuals.

5.2 Model behaviour with default parameters

We can analyse outcomes of a disease outbreak through estimating whether it operates in a subcritical, critical, or supercritical environment. Such estimates can be obtained through considering an expected number of child clusters per cluster, which would depend on the underlying parameters $(\beta, \gamma, p, \pi, \theta, \lambda)$. As part of the simulations, we start at a single household V_0, and go through the following process:

- Calculate the total rate of local infection, global infection, recovery and detection
- Generate waiting time until next event
- Choose an event from {birth, infection, recovery, confirmation}
- Choose which vertex to assign such event to
- Update the system and corresponding rates based on the event

After setting parameters to default values and running 200 simulations, the results presented in Figure 6 were obtained.

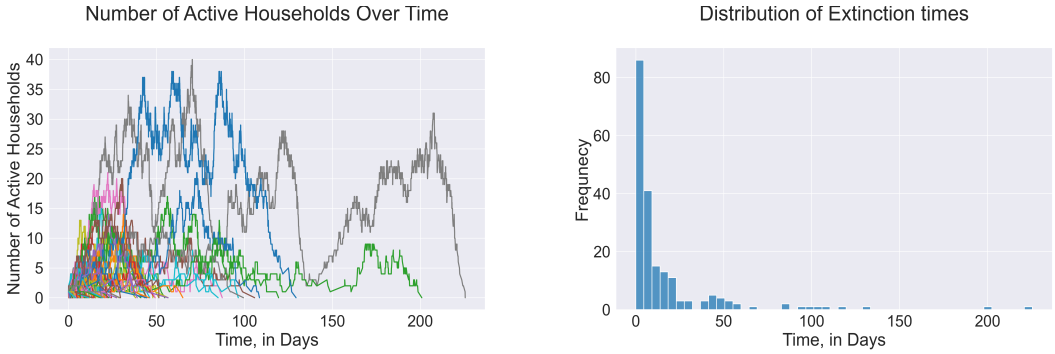


Figure 6: Number of active households in 200 simulations with default parameters

From the above, we notice that in order for an epidemic to spread, it needs to have benign conditions and relatively high transmissibility. Given the default parameters, none of the simulated 200 scenarios led to a disease surviving past one year, with majority of cases ceasing to exist within the first month. Such behaviour indicates that the expected number of child clusters per each cluster is likely to be below one, and we are likely to be in the subcritical regime.

To investigate such claim further, we need to estimate an expected number of untraceable edges generated by a typical cluster. However, this already presents two problems. Problem number one is that if we run a simulation and stop it at some predetermined time, call it $t = T$, then we end up with a large proportion of clusters which haven't yet been isolated, and hence have a potential to generate more child clusters. Hence, we cannot consider such clusters in our calculation, as otherwise an estimate is going to be negatively biased. Furthermore, if we run a simulation and then consider clusters which have already been isolated, we end up picking "unlucky" clusters which had shorter lifespan, and hence we will once again end up with a negative bias. Problem number two is that if we run our simulation until some predetermined time $t = T$ and then only consider the first few clusters in the process that have already been either isolated or fully recovered, we will end up having positive bias. Such bias comes from the fact that clusters that originated early in the process are likely to give births early, and hence have more children. It is therefore important to keep it in mind when working with simulations that require us to stop the process at a certain point (e.g. when we start getting into supercritical regime).

For the default parameters, however, we are able to run 200 simulations without having to stop a process at some predetermined time. We therefore arrive at the following statistics from a batch of simulations:

Number of simulations	Total number of clusters	Average children	Average size
200	704	0.716	1.051

Table 1: Cluster statistics with default parameter values

Notice that the average number of child clusters is below 1, which is consistent with the results presented in Figure 6. Both results point towards the default parameters creating

a subcritical regime environment, which would lead to an extinction of a disease with probability equal to 1. Furthermore, the average size of a cluster is low, at 1.051. Such result is not surprising as at the beginning of the process we have only one infected individual that is trying to infect others, while we additionally have recovery and confirmation events acting on the same individual. In the case of an initial household consisting of five individuals, for example, the probability that the first event would lead to extinction is $\frac{1}{3}$. Furthermore, the probability of an edge being traceable is 0.06, which means that a substantial number of clusters will be able to recover or get detected by the time one cluster will attach a traceable household. This results in most clusters consisting of a single node.

5.3 Transmissibility

As the next step, it is interesting to vary the local and global rate parameters, as those would correspond to different types of diseases and connectivity of people beyond household settings. We then estimate the resulting average number of child clusters for each of the scenarios. Since in this setting we will cover cases where we have to stop simulation at a predetermined time, we have to keep in mind that such estimates are going to be biased. In order to avoid a large negative bias, we will run each simulation until the first few clusters become inactive, and then only consider those first few clusters. On the other hand, this means that we end up with a positive bias, however, such bias is smaller compared to the first case. The results of such simulations are presented in Figure 7.

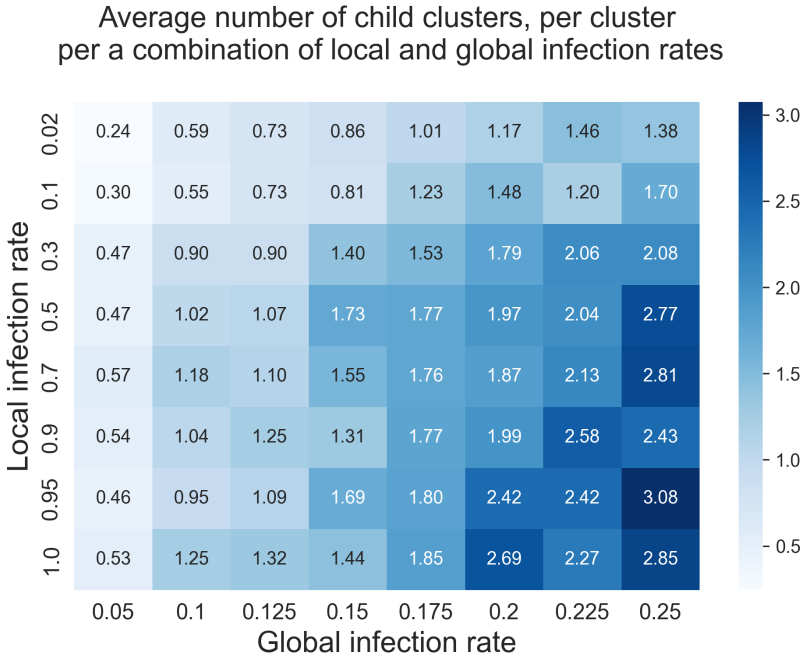


Figure 7: Heatmap of the average number of child clusters that a typical cluster generates, with 200 clusters used in a calculation of the average, per combination of infection rate parameters. The other parameters are kept at their default values.

Overall, we notice that as infection rates increase, it results in a higher number of child

clusters per typical cluster, which is in line with expectations. For the global infection rate of 0.05, all of the local infection rates lead to an estimated subcritical regime, even given the fact that such estimate is positively biased. Such scenarios are likely to arise during strict lockdowns, where local infection rate may still be high, however, without infecting enough new households the disease is bound to die out. This result highlights an effectiveness of the strategy where population has to stay within their respective households, and only a small number of critical global interactions are allowed. We also highlight that when $\beta = 0.05$, a large change in the local infection rate, e.g. from 0.1 to 1, results only in a modest increase in the estimated average number of child clusters, from 0.3 to 0.53. This, in turn, highlights the effectiveness of lockdowns for a wide range of diseases.

As we move towards higher global infection rates, we start obtaining results which are closer to a critical regime. For example, for $\beta = 0.1$, and $\lambda = 0.5$ our estimate reaches the value of 1.02, and increases further as we increase either global or local infection rates. This indicates that in the absence of strict lockdowns, even if community has a low connectivity between members of different households, it is still possible for a disease to reach a critical state if its transmissibility is high enough. Hence, for certain highly infectious diseases one would require a strict policy of isolation in order to stop the disease from spreading. On the other hand, even if local infection rate stays at a low level, such as 0.1, one can still end up with critical or supercritical regime if global infection rate is high. Such scenario can arise during a public holiday, festivals or other events in the area which lead to crowded spaces. We note, however, that the case of a local infection rate being 0.02, while global rate is above 0.2 is quite unlikely, as an infected person will have to stay in close contact with a very large number of individuals. Nevertheless, we still include these combinations of parameters for completeness.

As mentioned previously, however, such results retain positive bias, so in order to validate the data presented in the heatmap we additionally consider trends in the number of active households through time. For this, we analyse different levels of global infection rate while the local infection rate is fixed at 0.3. Such analysis is interesting, as its results correspond to the impact of connectivity between individuals beyond households on how quickly a disease spreads. The results, which clearly support the data presented in the heatmap, can be found in Figure 8 and Table 2.

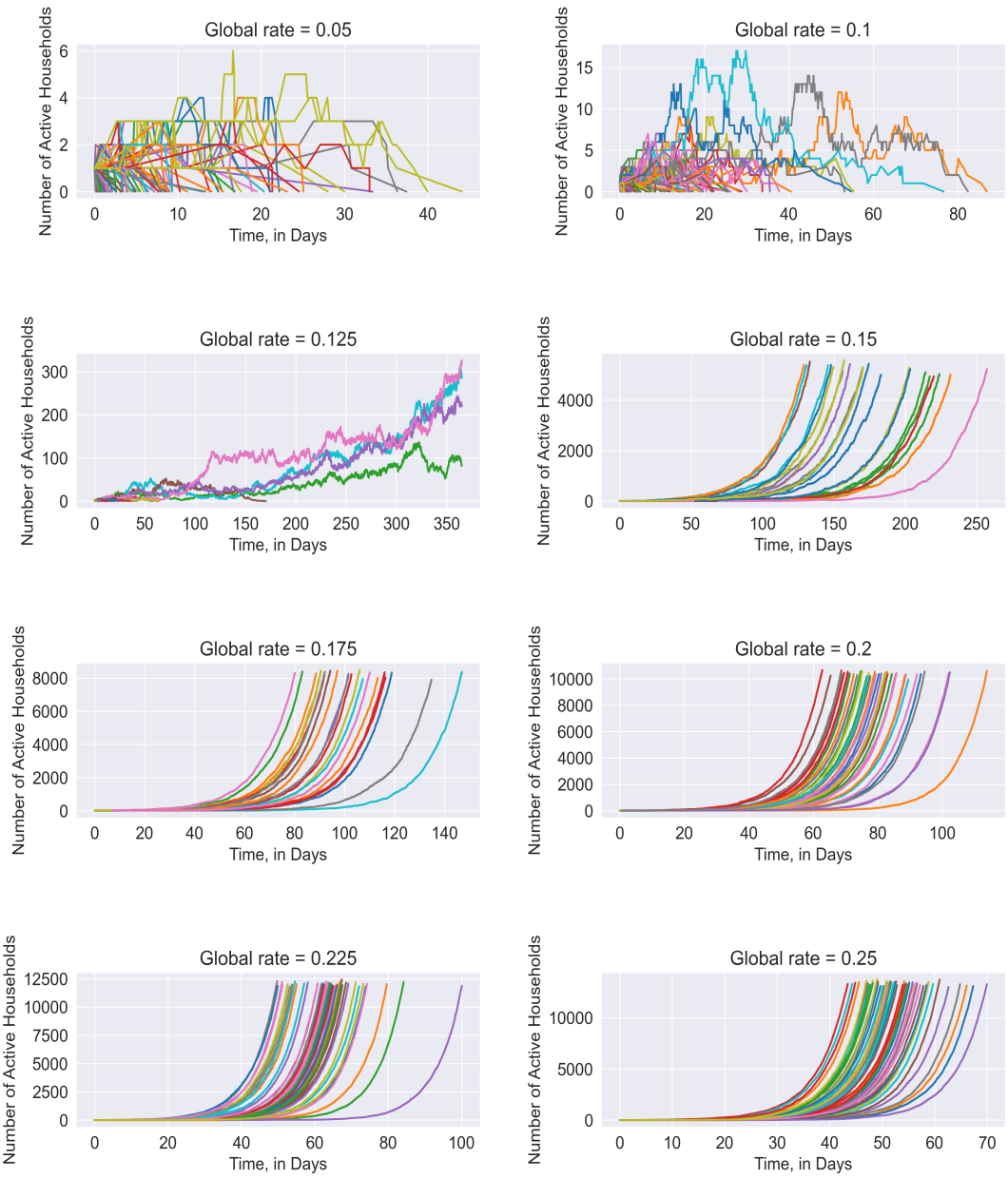


Figure 8: Simulations of the number of active nodes through time, for different levels of global infection rate, β. Local infection rate, λ, is set to 0.3, and other parameters – to their default values. Each plot includes 100 simulations.

Global rate	0.05	0.1	0.125	0.15	0.175	0.2	0.225	0.25
Survival past 1 year	0	0	4	22	22	42	44	52

Table 2: Survival cases, per 100 simulations, for different levels of global infection rate

As can be seen from Figure 8, the plot corresponding to the lowest global infection rate leads to a disease becoming extinct in every case out of 100 simulations prior to that disease hitting two months mark in time. This, once again, highlights an effectiveness

of lockdowns when trying to hault a disease from spreading. We further notice that the maximum number of households that were active at any single point in time is 6, which was reached in only one simulation. Most of the other cases were stopped before the root vertex managed to infect any other household.

As we move to global rates beyond 0.1 we start seeing cases where disease would survive past one year. In the case of $\beta = 0.125$, we end up with three simulations that resulted in over 200 households being infected within one year. This demonstrates that given benign conditions, even if a neighbourhood has low connectivity between its residents, there is a possibility for a disease to spread.

As we move past the global infection rate of 0.15, we start seeing a substantial proportion of cases behaving in an exponential explosion manner. For $\beta = 0.2$, for example, 42% of cases managed to survive past one year, with the total number of infected households going beyond the maximum number specified as a stopping criteria in a simulation. Such stopping criteria was implemented due to time and computational resources constraints.

Finally, as we reach the maximum global infection rate within our range, i.e. $\beta = 0.25$, we see more than a half of cases surviving past one year, with an exponential increase in the number of infected households trend being observed. These are types of cases in which one would desperately require additional measures, such as tracing and isolation, to bring the disease under control.

In this section, we relied on simulations of the underlying nodes and clusters, and hence each simulation had a high variance with regards to its outcome. Furthermore, as previously mentioned, an estimate of an expected number of child clusters generated by a typical cluster (which corresponds to BRN) would still retain bias after taking an average of simulations results. In the next section, we introduce another way one can arrive at an estimate of the Basic Reproduction Number. Such estimate is going to have a smaller variance and will be unbiased, and hence will allow us to arrive at a more reliable estimate of an outcome of an outbreak.

6 Basic Reproduction Number (R_0)

Basic Reproduction Number is a commonly used metric which has been defined in section 2, however, we will use a slightly different notation. Since we are analysing the dynamics of the system from the perspective of evolution of clusters as a Crump-Mode-Jagers process, we define R_0 in the following way:

Definition 4 (Basic Reproduction Number (BRN or R_0))**.** Basic reproduction number is the number of secondary infections generated by a single cluster prior to its recovery or isolation, where such infection is made to an individual in a household outside of the current cluster, with untraceable edge. In our setting, such infection can be made to any individual out of an infinite pool of susceptible individuals.

6.1 Stochastic Representation

In order to obtain an unbiased estimate of the Basic Reproduction Number, we write out stochastic differential equations of the system below, and use those to derive R_0 as an expected number of untraceable global contaminations. In this section, we continue working with clusters as a way to analyse the system, and, once again, consider "birth" to be a generation of an untraceable edge by a cluster.

Define the following quantities:

- Let $C(t)$ be the number of households at time t present in a typical cluster
- Let $I(t)$ be the number of infected individuals in a cluster at time t
- Let $I_k(t)$ be the number of infected individuals in household k that is present in the cluster
- Let $R_k(t)$ be the number of recovered individuals at time t in household k
- Let $D(t)$ represent detection, where 0 indicates that the cluster has not yet been detected and 1 - that it has been detected

Then, we have the following:

$$C(t) = \begin{cases} C(0) + \text{Poisson}\{\int_0^t \beta p I(s) ds\}, & \text{if } D(t) = 0 \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

$$R_k(t) = 0 + \text{Poisson}\{\int_0^t \gamma I_k(s) ds\} \tag{26}$$

$$I(t) = \begin{cases} \sum_{k=1}^{C(t)} \left[\sum_{n=1}^{\infty} \pi_n \left[I_k(0) - R_k(s) + \text{Poisson}(\phi) | h_k = n \right] \right], & \text{if } D(t) = 0 \\ 0, & \text{otherwise} \end{cases}$$

where ϕ is an infection rate that can be expressed as follows:

$$\int_0^t \lambda \times I_k(s) \frac{h_k - I_k(s) - R_k(s)}{h_k - 1} ds, \tag{27}$$

In this case, $D(t)$ is an indicator of whether detection within a cluster has occurred, where detection happens at rate $\theta I(t)$. If $D(t)$ becomes 1, then $I(t)$ becomes 0 for all future values of t.

This leads to the following Basic Reproduction Number:

$$R_0 = \beta(1-p) \times E[\int_0^\infty I(t)dt \;\mid\; I(0) = 1,\; I_k(0) = 1\;\forall k,\;\; C(0) = 1, R_k(0) = 0\;\forall k] \tag{28}$$

In order to calculate R_0 we rely on simulations. More concretely, we create a discrete approximation of the above stochastic representation with time increment of $dt = 0.01$, and then follow the below steps:

- Initialise parameters for $\pi, \beta, \gamma, \lambda, \theta, p$
- Create the following loop to simulate 100 clusters:
  - Initialise quantities of interest:
    - Set $C(0) = 1$
    - Set $R_1(0) = 0$
    - Set $I(0) = 1$
    - Set $D(t) = 0$
    - Set $I_1(0) = [1]$
    - Set $R_1(0) = [0]$
  - Grow such cluster until the number of infected individuals for the whole cluster, $I(t)$, is not set to 0. If $I(t)$ is equal to zero, then break the loop and start simulating the next cluster. $I(t)$ can become zero either due to a detection event, or due to all individuals recovering.
- Once 100 clusters are created, use Monte Carlo estimate for the expectation of the integral in calculation of BRN value
- Record one BRN value
- Repeat previous steps to create 100 BRN values

Given the default parameters, we arrive at the following estimate of R_0 for the default case:

Mean	Confidence interval	Min	Max
0.934	[0.904 , 0.963]	0.916	1.029

Table 3: Basic Reproduction Number statistics obtained from 100 clusters per simulation, over 100 simulations

As a result, using stochastic differential equations, we are able to arrive at an estimate of BRN with a relatively low standard deviation of 0.015. The obtained value of 0.934 highlights that we are likely to be in subcritical regime when considering default parameter

values, which is consistent with Figure 6. Furthermore, the heatmap in Figure 8 includes a value of 1.02 for the default rate parameters. Since we mentioned in section 5.3 that such estimate presented in the heatmap would be positively biased, a lower estimate of 0.934 is in line with our expectations.

6.2 Public health measures

In our model, the two opposing forces to the spread of a disease is recovery and isolation. It is therefore of interest to analyse their effects on the speed at which a disease spreads (or dies out). In this section, all R_0 values have been estimated using the discrete approximation of the process presented in section 6.1.

6.2.1 Detection rate

One of the ways to control the spread of epidemics is establishing measures that help individuals identify whether they have a disease in a timely manner. As an example, one can provide access to free health check ups or tests that would help an individual establish whether he may be suffering from a disease. In addition to that, one can impose isolation measures where an infected individual would be instructed to stay at home. Resources are not unlimited, however, and such strategies can become prohibitively expensive when applied to the whole population. For this reason, we analyse the levels of detection and isolation for which R_0 drops below 1.

Detection rate (θ)	Mean	Confidence interval	Min	Max
0.0	2.038	[1.988 , 2.088]	1.983	2.237
0.1	0.610	[0.582 , 0.638]	0.600	0.701
0.2	0.357	[0.347 , 0.367]	0.328	0.370
0.5	0.161	[0.155 , 0.167]	0.153	0.163
1.0	0.086	[0.084 , 0.088]	0.078	0.087

Table 4: Basic Reproduction Number statistics obtained from 100 clusters per simulation

As can be seen from Table 4, having no detection and isolation efforts results in R_0 of approximately 2, which is within the supercritical case scenario. Such result is unsurprising since the only opposing force to infections is recovery, which stays at 0.15 – considerably lower than the local infection rate, which is at 0.5. Once the number of infected individuals in a single household increases above one, this further leads to global infection rate being 0.2 or above – again, higher than the rate of recovery. This is likely to lead to a rapid increase in the number of cases.

As we move towards a slightly higher detection rate of 0.1 we can see a large drop in R_0 from 2.038 to 0.610, i.e. an estimated subcritical regime. This emphasises that detection and isolation, if implemented from the early stages of spread of a disease, can have a significant impact on the total number of people infected. This demonstrates that even a simple strategy that allows testing to be conducted for a subsection of the population, such as for the highest risk individuals, can result in establishing control over how quickly the disease spreads.

Finally, as detection rates approach the value of 1, the estimate of R_0 drops to a level below 0.1, which would lead to a rapid decrease in the number of infected individuals, provided that recovery rate is at approximately default level.

6.2.2 Traceability and Detection for highly transmissible diseases

One of the measures that can be implemented to help control the speed at which a disease is spreading is contact tracing. This can aid with notifying those at risk of getting an infection, and, if coupled with detection and isolation strategy, will instruct a timely isolation of an individual, which would decrease the number of people she would have infected otherwise. We would expect such measures to be in place whenever a disease is spreading rapidly through the population, and hence this section concentrates on highly transmissible diseases, with λ set to 1, and β – to 0.5. For this scenario, the heatmap in Figure 9 was obtained.

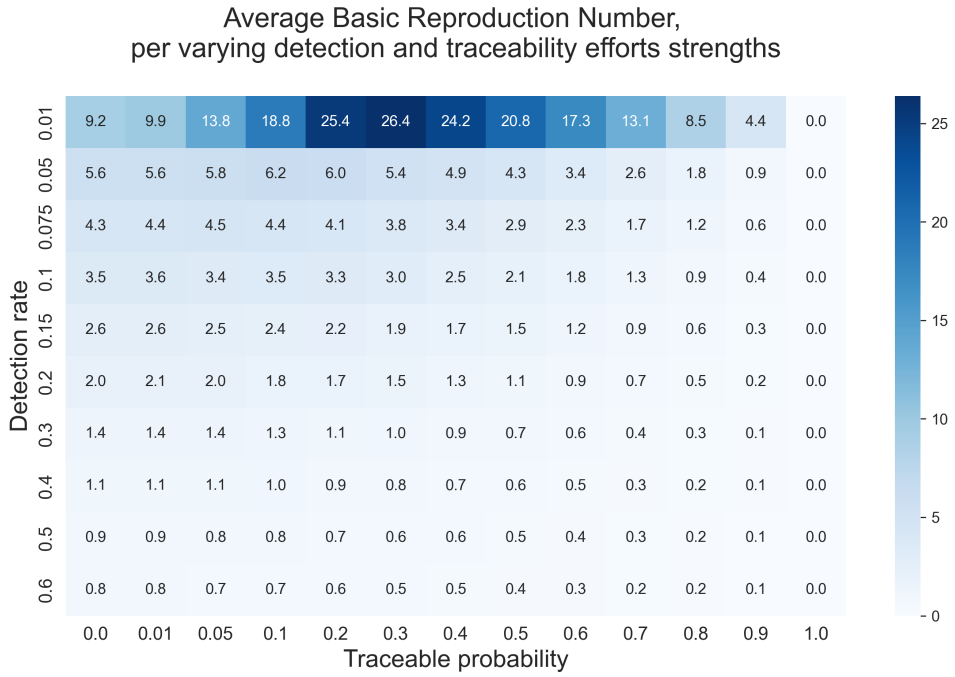


Figure 9: Estimation of the BRN for varying detection rate and traceable probability parameters

First, we notice that even a small increase in detection rate results in a large drop in the estimated R_0. If, for traceable probability of 0.01 and detection rate of 0.01, we have $R_0 = 9.9$, then once we increase the detection rate to 0.1, an R_0 drops to 3.6. Once detection rate reaches 0.5, even for a small traceable probability such as 0.01, we get into the estimated subcritical regime. It is worth noting, however, that detection rate of 0.5 or above may be difficult to obtain, as some people may not show symptoms, or would choose to not be tested. For this reason, it is important to couple detection together with

an increase in contact tracing strength, which would result in obtaining subcritical regime at lower levels of detection.

Secondly, we notice that for a low detection rate of 0.01, R_0 is non-monotone with respect to traceable probability, and it obtains the highest values at p being between 20-40%. This happens due to our definition of R_0. As a reminder, we define R_0 to be the average number of new global infections with untraceable edges generated by a cluster. In the case of low traceable probability and low detection rate we tend to have most clusters with $C(t) = 1$. Each of those clusters can therefore be detected at approximately the same rate. On the other hand, as we move towards higher traceable probability we start seeing larger clusters appearing. However, those larger clusters will now have higher detection rate and hence are more likely to be isolated. We therefore end up with those clusters getting isolated, which have "used up" their birth on traceable children which do not count towards the calculation of R_0. As we move towards traceable probability beyond 40%, however, we start seeing clusters of larger sizes, where detection leads to isolation of a substantial proportion of active nodes. This results in a decrease in the estimated R_0 values for traceable probabilities of 50% and above.

6.2.3 Recovery rate

It is also of interest to analyse the effect of recovery rates on the BRN. Let's assume there are a number of ways in which individuals can decrease the length of an illness. As an example, for some diseases, vaccinations tend to ease the effects of an illness as well as decrease the length of time that a person can transmit the disease to others. Access to medication can sometimes also decrease the length of recovery.

In order to analyse the effect of recovery on the BRN value, we first consider the model introduced by Jean Bertoin in [2] and described in section 3.2. On one hand side, Bertoin's model doesn't incorporate any societal structure, which makes it more difficult to compare it with the model proposed in our work. However, a benefit of his model includes a possibility of calculating R_0 explicitly in the case of no recovery, which lets us verify the value produced by our simulations. Hence, to allow for a fair comparison, we first set all households in our model to consist of a single individual, then estimate R_0 using simulations and compare it to the value of R_0 that is calculated explicitly. After verifying the result, we then additionally incorporate recovery dynamic and estimate the corresponding R_0 using simulations.

We run the model with the following parameters:

- All household sizes are set to 1 with probability 1
- $\lambda = 0$
- $\beta = 0.5$
- $\theta = 0.05$
- $p = 0.06$,

and recovery rate γ is the parameter being varied.

Then, the value of R_0 for the case of no recovery can be calculated explicitly, as in [2], via the following:

$$R_0 = \frac{(1-p)\beta}{\theta} \tag{29}$$

The results of such simulations are presented in Table 5.

	R_0 estimate	R_0 confidence interval
Explicit calculation for $\gamma = 0$	9.400	-
$\gamma = 0.0$	9.201	[8.969 , 9.433]
$\gamma = 0.1$	3.670	[3.596 , 3.744]
$\gamma = 0.2$	2.094	[2.036 , 2.152]
$\gamma = 0.5$	0.892	[0.852 , 0.932]
$\gamma = 1.0$	0.454	[0.43 , 0.478]

Table 5: An explicit calculation of R_0 for the model in [2] and an estimated R_0 with varying recovery rates, for $\lambda = 0, \beta = 0.5, \theta = 0.05, p = 0.06$, and all household sizes being set to 1

As can be seen from Table 5, the explicitly calculated value of R_0 falls within the confidence interval obtained from simulations, which serves as a positive check. Furthermore, we notice that the rate of recovery has a substantial effect on the average number of new child clusters generated by a typical cluster. If we consider $\gamma = 0.1$, which corresponds to an average length of illness of ten days, we obtain an R_0 of 3.670. However, as we decrease the length of illness to an average value of five days, the corresponding R_0 falls by 43% to 2.094. This shows that even if recovery rate is lower compared to the rate at which infection spreads, it can still have a significant impact on the outcome. Such results demonstrate an importance of vaccinations and appropriate medical care.

It is important to note, however, that this model was run with all households having a single individual, which, as a result, ignores household structure. Since household structure will have an additional effect on how an epidemic spreads, more analysis can be done to estimate an effect of recovery rates on resulting BRN values when taking societal structure into consideration.

7 Distribution of cluster sizes at isolation

One other quantity of interest when analysing preventative measures against spread of a disease is the distribution of cluster sizes at an instant prior to their isolation. Being able to predict an average number of isolated households whenever a single detection event occurs could be useful in estimating the total number of individuals in a population that could be isolated throughout a specific time period, provided a predetermined rate of detection and traceability.

Define a random variable X denoting the size of a typical cluster at an instant prior to its isolation. In [2], the author notes that X would follow a geometric distribution. However, due to household structure, we cannot expect X to be geometrically distributed in our model. Nevertheless, it is interesting to assess how its distribution would differ compared to the case with no household structure and no recovery.

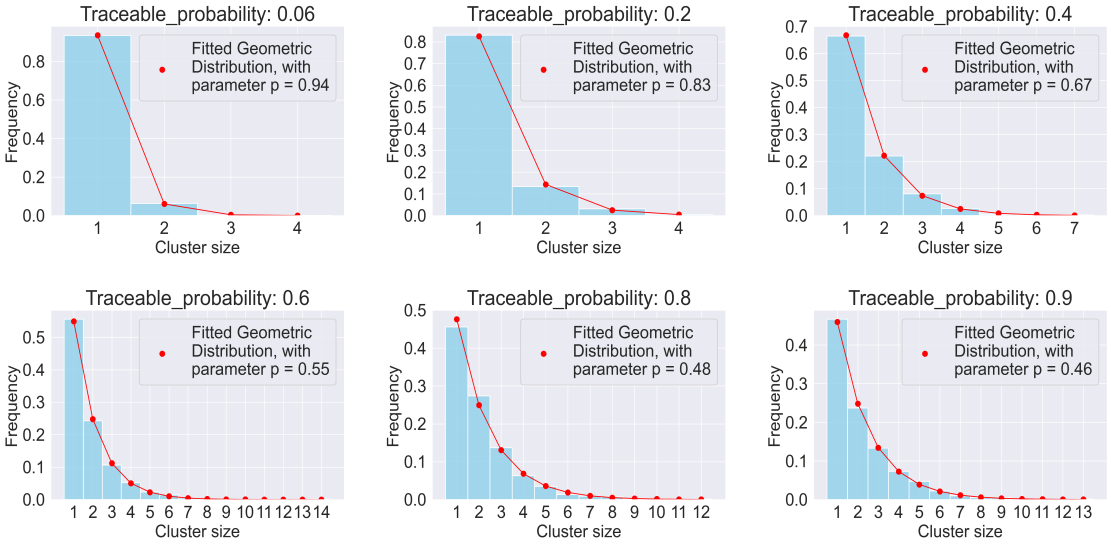


Figure 10: Simulated distribution of X, for different levels of traceable probability p, and other parameters at default values. Simulations of 1000 clusters per plot.

The plots in Figure 10 demonstrate the simulated distribution of X for varying values of p. Even though the data seem to follow a geometric distribution when assessed visually, a p-value of 3.15×10^{-161} was obtained when running Kolmogorov-Smirnov test. This indicates that we reject the null hypothesis of data coming from the geometric distribution.

To assess where the differences come from, we plot a semi-log plot in Figure 11, where y-axis has a log-scale, and x-axis has a linear scale. The probability mass function of the geometric distribution is $P(X = x) = (1-p)^{x-1}p$, which, once we take a logarithm of both sides, becomes $\log(P(X = x)) = (x-1)\log(1-p) + \log(p)$. Hence, if the data was geometrically distributed, the log of probabilities would be approximately linear in x. Such plot is visualised in Figure 11 with $p = 0.3$.

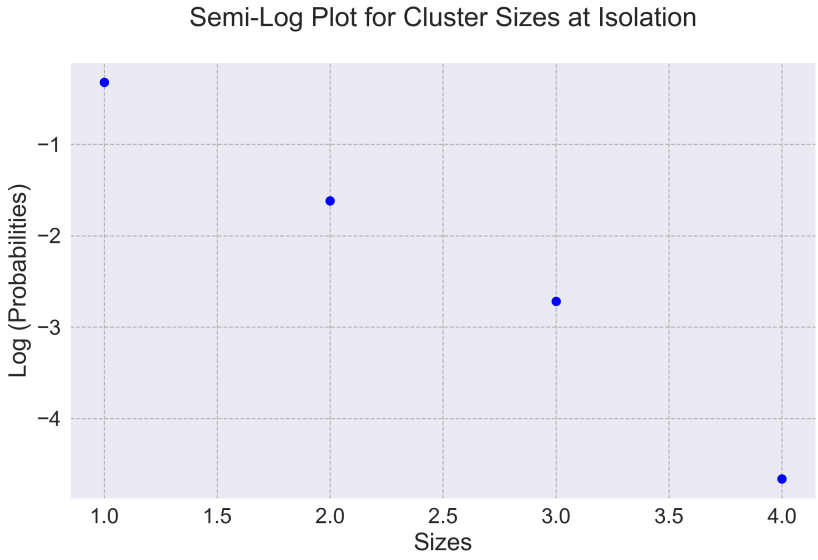


Figure 11: Semi-log plot of the cluster sizes at isolation data with default parameters and $p = 0.3$.

As can be seen from Figure 11, the relationship seems linear for cluster sizes up to $C(t) = 3$. At $C(t) = 4$, however, we observe less values than we would if the data was coming from the geometric distribution. This result is in line with our expectations, indicating that incorporating household structure and recovery leads to smaller clusters being isolated. However, we also note that the effect of recovery on the distribution of X is not straightforward, as each recovery inside a cluster will decrease the rate of detection for that cluster. Such relationship is therefore dependent on the values set for parameters in a simulation.

8 Numerical applications: testing and isolation strategy in London

Given that the latest pandemic resulted in a large amount of information available about the disease, this section concentrates on the case of COVID-19 in London. We consider the transmissibility rates and other characteristics obtained about the disease during the beginning of its outbreak, as well as adjust the distribution of individuals per household parameter based on London's demographics. We then infer the strength of testing and isolation that were required in order to stop the spread of the disease in a reasonable time.

Let us start with London's demographics, and adjust our model's parameter accordingly. As of 2020, which was the year COVID-19 started spreading in the UK, the average household size in London was 2.68 people per household. The proportions, as per [22], were as follows:

Number of people per household	Percentage of households
One person	23.5%
Two people	29.0%
Three people	18.2%
Four people	19.4%
Five people	6.4%
Six people	2.5%
Seven or more people	1.0%

Table 6: London Households by size, data from *Statista* (2023)

Given the observed trend in percentages of going from four to seven or more people per household, we will use 0.6% for seven people households, 0.3% for eight people households, and 0.1% for nine people households when running the simulation.

Next, we move on to infection and recovery rates, where we consider empirical research from 2020, as we want to concentrate on the case of stopping the spread of the disease early on before new mutations occur. Furthermore, we concentrate on studies that relate specifically to the UK, as different countries will have different traditions and community structures, and hence their global infection rates will vary. Given the above, we chose to rely on [18], which was published in August 2020 - half a year after the first COVID-19 case was detected in the UK. Based on the data provided in the paper, we set our local infection rate to 0.47 and global infection rate to 0.12. Finally, given data from [8] we take the recovery rate to be 0.2.

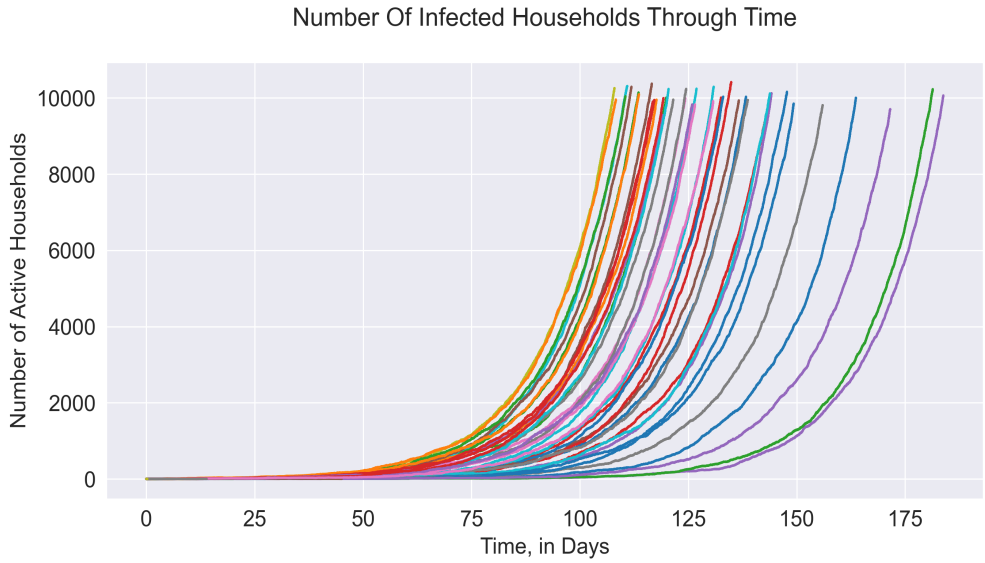


Figure 12: Number of infected households through time, based on 100 simulations, with no intervention ($\theta = 0, p = 0$) and $\beta, \lambda, \pi, \gamma$ parameters based on London's demographics and COVID-19 2020 empirical data.

Initially, a simulation of a spread of COVID-19 was run to estimate the probability of a rapid increase in the number of cases. The case with no intervention ($\theta = 0, p = 0$) led to 39 out of 100 simulations resulting in an exponential trend for such cases being observed, as can be seen from Figure 12. In such cases, one would require additional measures to slow down the spread of the disease and make sure the region has enough resources to cope with a wave of new patients.

As a way to measure the level of intervention required, estimates of BRN and extinction probabilities for varying levels of contact tracing and isolation have been obtained. Such results are presented in Figure 13, where we present an average of unbiased estimates of R_0, per combinations of parameters.

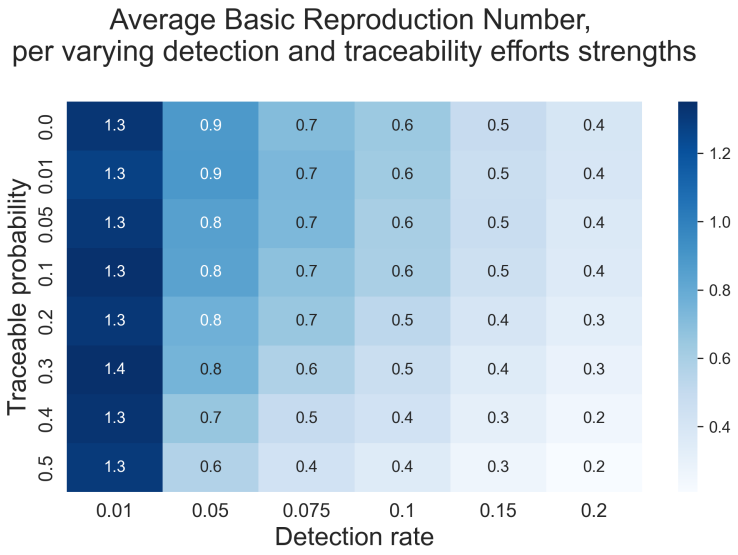


Figure 13: Effect of detection and tracing on BRN for COVID-19 data in London.

Interestingly enough, from Figure 13 we can observe that even a small detection rate of 0.05 results in an estimated R_0 dropping from the supercritical to subcritical case, for any contact tracing level. One has to keep in mind, however, that such results are obtained based on an isolation strategy implemented from the very beginning of the disease appearing in London. In reality, there are likely to be at least a few months passing before an intervention can be implemented. Hence, as a potential improvement to the described model, one can further consider cases of intervention efforts, including traceability, starting to take effect only after a certain number of individuals have been infected. Nevertheless, the above result is useful as it highlights how an intervention, even if it affects a small proportion of the population, can have a significant impact on the outbreak outcome.

Furthermore, when estimating probability of the disease extinction within the first year at different levels of intervention, the results presented in Table 7 were obtained:

	Traceable probability							
Confirmation rate	0.00	0.01	0.05	0.10	0.20	0.30	0.40	0.50
0.00	0.61	0.76	0.66	0.69	0.73	0.67	0.69	0.71
0.01	0.80	0.75	0.81	0.82	0.87	0.78	0.82	0.80
0.05	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.075	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 7: Estimates of probability of extinction within one year, based on 100 simulations, for combinations of detection rate and traceable probability, with $\lambda = 0.47$, $\beta = 0.12$, $\gamma = 0.2$ and size-biased household distribution based on London's demographics in 2020.

From Table 7 we can observe that for all cases where confirmation rate is at a level of 0.05 or above, none of the simulated scenarios led to even a single survival of the disease

past one year. This result is in line with Figure 13, where we have seen that a detection rate of 0.05 or above leads to R_0 dropping below 1. Furthermore, we can see that even for such a small confirmation rate as 0.01, over 70% of scenarios lead to a disease extinction within one year for all traceable probability levels. The main consequence of the above results lies in an importance of early intervention, even if such intervention affects a small proportion of the population. The data therefore shows that the case of COVID-19 in London, in 2020, could have been brought to stable levels in the same year if the prevention strategies, such as providing advice to stay at home in case of symptoms and availability of testing kits, were to be implemented early.

9 Model limitations

It is worth briefly discussing model limitations and potential improvements that can be made such that the model becomes more realistic. First, as mentioned in section 4, we only consider societal structure from a perspective of households. This means that we ignore other potential hyper-edges, such as the ones representing circles of colleagues or neighbourhoods. Given that people spend a large part of their day at work, this would be the next best variable to include. Additionally, we assumed that the local infection rate is constant through time. This means that the model doesn't take into account potential new mutations and their impact on the level of transmissibility. The model can be, however, easily adjusted to incorporate varying local infection rates, if such data becomes available. Finally, as mentioned in the previous section, it would also be interesting to assess whether delayed isolation and traceability efforts would result in a similarly strong effect in decreasing R_0.

10 Conclusion

Over the course of working with the proposed model and obtaining simulation results, one insight becomes apparent – an early intervention can have a significant impact on the outbreak outcome, even if such intervention is limited in its outreach. This insight is supported by significant drops in R_0 value, as well as drop in the number of disease cases surviving past one year in simulations, as soon as any preventative measure is introduced in the model. This, in turn, highlights the value of developing methods that would be able to detect rising cases of diseases early. Examples of such methods involve tools that would highlight rapid increases in cases related to any specific disease across a network of hospitals. Another method could involve raising awareness about importance of reporting cases that would aid with tracking of where and when such cases occur.

Our model relies on estimates of global and local infection rate, as well as recovery rate, and hence it becomes necessary to be able to get reliable estimates. Whether this, or any other model, is used for determining the appropriate levels of interventions, the early estimates of the characteristics of a disease become crucial. This means, in turn, that collection of appropriate data across affected locations becomes critical in establishing the correct response. It shouldn't be understated, therefore, that maintenance and accessibility of such data can save time and resources, as it can help with establishing measures in time, instead of relying on trial and error. Reliable data can also save time and money in case such measures are not needed and disease is likely to die out by itself with high probability.

In real world, it is impossible to capture every interaction and parameter that may affect the way that an outbreak spreads out. However, identifying the most important parameters can help us predict an evolution of an outbreak and hence get it under control. In our case, one of those parameters includes household structure. Incorporating societal structure made our model more complex, and hence more difficult to analyse, but it also made it more realistic and hence applicable. More work can be done with respect to incorporating circles of colleagues and friends, to better capture the non-random interactions between individuals. Our model takes the first step in this direction.

References

[1] Debasis Bagchi, Amitava Das, and Bernard William Downs, editors. *Viral, Parasitic, Bacterial, and Fungal Infections: Antimicrobial, Host Defense, and Therapeutic Strategies*. Academic Press, 1 edition, 2022.

[2] Jean Bertoin. A model for an epidemic with contact tracing and cluster isolation, and a detection paradox. arXiv: 2201.01924 [math.PR], 2022.

[3] Simon Cauchemez, Christl A. Donnelly, Carrie Reed, Azra C. Ghani, Christophe Fraser, Charlotte K. Kent, Lyn Finelli, and Neil M. Ferguson. Household transmission of 2009 pandemic influenza a (h1n1) virus in the united states. *The New England Journal of Medicine*, 361:2619–2627, December 2009. Accessed: 2023-07-29.

[4] European Commission. Mobile contact tracing apps in eu member states, 2023. Accessed: 2023-08-01.

[5] Moez Draief and Laurent Massoulié. Galton–watson branching processes. In *Epidemics and Rumours in Complex Networks*, London Mathematical Society Lecture Note Series, pages 7–18. Cambridge University Press, 2009.

[6] Raphaël Forien and Étienne Pardoux. Household epidemic models and mckean-vlasov poisson driven stochastic differential equations. arXiv: 1907.03001 [math.PR], 2022.

[7] World Economic Forum. The global risks report 2023 18th edition, 2023. Chapter: Global Risks 2023: Today's Crisis. Published on 11 January 2023. See Figure 1.1.

[8] Seran Hakki, Jie Zhou, Jakob Jonnerby, and et. al Singanayagam. Onset and window of sars-cov-2 infectiousness and temporal correlation with symptom onset: a prospective, longitudinal, community cohort study. *The Lancet Respiratory Medicine*, 10(11):P1061–1073, 2022.

[9] NHS Inform. Illnesses and conditions: Infections and poisoning, 2023. Section: Flu, Accessed: 2023-06-29.

[10] Peter Jagers and Olle Nerman. The growth and composition of branching populations. *Advances in Applied Probability*, 16.2:221–259, 1984. Accessed: 2023-07-20.

[11] M. Kendall, D. Tsallis, C. Wymant, A. Di Francia, Y. Balogun, X. Didelot, L. Ferretti, and C. Fraser. Epidemiological impacts of the nhs covid-19 app in england and wales throughout its first year. *Nat Commun*, 14(1):858, February 2023. PMID: 36813770; PMCID: PMC9947127.

[12] William Ogilvy Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A*, 115(772), 1927.

[13] Robert C. King, Pamela K. Mulligan, and William D. Stansfield. *A Dictionary of Genetics*. Oxford University Press, 8 edition, 2013.

[14] WC Koh, L Naing, L Chaw, MA Rosledzana, MF Alikhan, SA Jamaludin, F Amin, A Omar, A Shazli, M Griffith, R Pastore, and J Wong. What do we know about

sars-cov-2 transmission? a systematic review and meta-analysis of the secondary attack rate and associated risk factors. *PLoS One*, 15(10):e0240205, October 2020. PMID: 33031427; PMCID: PMC7544065.

[15] Amaury Lambert, Florian Simatos, and Bert Zwart. Scaling limits via excursion theory: Interplay between crump-mode-jagers branching processes and processor-sharing queues. *The Annals of Applied Probability*, 23(6):2357–2381, 2013.

[16] Jonathan Law and Elizabeth Martin, editors. *Concise Medical Dictionary*. Oxford University Press, 10 edition, 2020.

[17] Q. Li, X. Guan, P. Wu, X. Wang, and L. et. al Zhou. Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia. *N Engl J Med*, 382(13):1199–1207, 2020. Epub 2020 Jan 29.

[18] Jamie Lopez Bernal, Nikolaos Panagiotopoulos, Chloe Byers, et al. Transmission dynamics of covid-19 in household and community settings in the united kingdom. *Eurosurveillance*, 27(15), 2022.

[19] H. E. Stanley M. Dickison, S. Havlin. Epidemics on interconnected networks. arXiv: 1201.6339 [physics.soc-ph], 2012.

[20] World Health Organisation. Transmission of sars-cov-2: implications for infection prevention precautions, 7 2020.

[21] Miquel Porta. *A Dictionary of Epidemiology*. Oxford University Press, 6 edition, 2014. Current Online Version: 2016, eISBN: 9780199390069.

[22] Statista. Households by household size, regions of england and gb constituent countries, 2023. Accessed: 2023-08-15.

[23] Plamen Trayanov. Crump-mode-jagers branching process: A numerical approach. *Branching Processes and Their Applications*, 219:167–182, 2016. Accessed: 2023-07-27.

[24] Chenlin Gu Vincent Bansaye and Linglong Yuan. A growth-fragmentation-isolation process on random recursive trees and contact tracing. arXiv: 2109.05760 [math.PR], 2022.

[25] Wikipedia. Galton–watson process, 2023. Accessed: 2023-08-10.

[26] C. Wymant, L. Ferretti, and D. et al. Tsallis. The epidemiological impact of the nhs covid-19 app. *Nature*, 594:408–412, May 2021.

Code Appendix

The Python code below only includes the main functions defined for simulations. Further code that was used to generate Figures and Tables in this work can be found in the following repository: github_repository_link