A model of COVID-19 propagation based on a gamma subordinated negative binomial branching process

Jérôme Levesque*, David W. Maybury†, and R. H. A. David Shaw
Public Services and Procurement Canada
270 Albert Street, Ottawa ON K1P 6N7
and
Public Health Agency of Canada
130 Colonnade Road
Ottawa, ON K1A 0K9

August 27, 2020

Abstract

We build a parsimonious Crump-Mode-Jagers continuous time branching process of COVID-19 propagation based on a negative binomial process subordinated by a gamma subordinator. By focusing on the stochastic nature of the process in small populations, our model provides decision making insight into mitigation strategies as an outbreak begins. Our model accommodates contact tracing and isolation, allowing for comparisons between different types of intervention. We emphasize a physical interpretation of the disease propagation throughout which affords analytical results for comparison to simulations. Our model provides a basis for decision makers to understand the likely trade-offs and consequences between alternative outbreak mitigation strategies particularly in office environments and confined work-spaces. Using our model with Bayesian hierarchical techniques, we provide US county level inferences of the reproduction number from cumulative case count data over July and August of this year.

1 Introduction

As of June 20, 2020, there have been more than 8 million confirmed global cases of COVID-19, a respiratory illness caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Early indications suggest a case/infection fatality rate of between 0.5% to 2% [1, 2, 3, 4] with poor prognosis strongly dependent on comorbidity

*Corresponding author: Jerome.Levesque@tpsgc-pwgsc.gc.ca
†Corresponding author: David.Maybury@tpsgc-pwgsc.gc.ca

factors such as advanced age, diabetes, and other poor health conditions [5]. The Centers for Disease Control and Prevention in the United States gives an overall current symptomatic case fatality ratio of 0.4% [6] while studies involving seroprevalence indicate a median infection fatality rate of 0.25% [7]. Canada has seen over 100,000 cases and the entire world has engaged in costly outbreak mitigation strategies to prevent excess deaths.

Governments around the world have focused on controlling COVID-19 outbreaks primarily by reducing direct human-to-human contact through varying degrees of society-wide lock-downs and strong social distancing measures. By limiting the opportunity for infectious contacts, the hope is that the infection rate will remain low enough to prevent medical support systems from becoming overwhelmed while also reducing the effective reproduction number of the disease. Evidence suggests that government lock-down strategies are having a positive effect [8], but those strategies may also become prohibitively expensive in the not too distant future. An alternative outbreak controlling strategy to lock-downs is contact tracing with isolation. In this strategy, health authorities trace the human-to-human contacts of an infected person and isolate those contacts who are at risk having become infected. If the probability of isolating potentially infected contacts is high and the time to isolation is sufficiently short, contact tracing with isolation may offer better cost benefit performance relative to lock-downs in keeping society safe [9].

Modelling the spread of infectious diseases falls into two broad classes [10]: deterministic modelling, which captures the thermodynamic limit and large scale behaviour of the underlying epidemiological phenomenon, and stochastic modelling, which describes the microscopic statistical nature of the generative process. Traditional compartmental models (e.g., SIRD), usually expressed as a set of coupled ordinary differential equations, fall into the first class while branching processes, in which each infected individual randomly generates "offspring", belong to the second class. In this paper, we focus on a stochastic formulation of COVID-19 following Hellewell et al. [11].

In [11], the authors develop a branching process to model contact tracing with isolation strategies. The model uses a negative binomial distribution to generate secondary cases produced by an infected individual with new infections assigned a time of infection through draws from a serial interval distribution. By truncating the serial interval distribution through isolation events, the authors show that in most of their scenarios contact tracing and case isolation is enough to control a new outbreak of COVID-19 within 3 months.

While the construction in [11] provides a rich base for numerical simulations, to gain further insight, we extend the model to a fully continuous time setting which provides us with a complete generative model, including expressions for the generating and characteristic functions. Furthermore, each part of our model has a direct physical interpretation of the underlying disease propagation mechanism. The model balances fidelity and parsimony so that the model can 1) be calibrated to data relatively easily, 2) provide semi-analytic tractability that allows for trade-off analysis between different mitigation strategies 3) generate realistic simulated sample paths for comparing interventions. Our code is available as an R package.

2 The model

In this paper, we build a Crump-Mode-Jagers (CMJ) branching process model through a subordinated Lévy process. CMJ constructions contain the triple of random processes $(\lambda_x, \xi_x(\cdot), \chi_x(\cdot))$ loosely defined as,

- λ_x is a random variable that denotes an infected person's communicable period;
- $\xi(t) = \#\{k : \sigma(\omega, k) \leq t\}$ counts the number of infected people over event space $\Omega(\omega)$ in time t. $\xi_x(t - \sigma_x)$ denotes the random number of infected people created by an infected person at every moment of her communicable period over the interval $[\sigma_x, t)$; $\xi(t - \sigma_x) = 0$ if $t - \sigma_x < 0$; and
- $\chi_x(t - \sigma_x)$ is a random characteristic of the infected person within the interval $[\sigma_x, t)$; $\chi(t - \sigma_x) = 0$ if $t - \sigma_x < 0$. (E.g., $\chi(t) = \mathbb{I}\{t \in [0, \lambda)\}$ is the number of infectious existing at moment t).

Our model generates infections from an infected individual through a compound Poisson process where the event times represent transmission events (σ_x). We imagine that an individual is infectious from the moment she becomes infected. The number of new infections at each transmission event is a draw from the logarithmic distribution [12] ($\xi(t)$) and consequently, the resulting generative process is the negative binomial process (see, for example, Quenouille [13]). The stochastic counting processes remains "on" during the communicable period and then shuts "off" at the end—that is, the communicable period is the random lifetime (λ_x) in the CMJ language. We model the communicable period as a gamma distributed random variable, $\Gamma(a, b)$, with mean $\bar{t} = a/b$. By subordinating our resulting negative binomial process with a gamma process for the communicable period, we arrive at our model of COVID-19 propagation—a gamma negative binomial branching process (GNBBP). (For details on subordinated Lévy processes, see [14].)

2.1 Construction details

We model the propagation of COVID-19 by assuming that people become infectious immediately after contracting the virus and that they can infect others throughout the duration of their communicable period. We assume the population is homogeneous and that each new infected individual has the same statistical properties as previously infected people. Specifically, we assume that an infected person infects $Q(t)$ other people during time interval $[0, t]$ according to a compound Poisson process,

$$Q(t) = \sum_{i=1}^{N(t)} Y_i, \tag{1}$$

where the number of infectious events, $N(t)$, follow a Poisson counting process with arrival rate λ, and Y_i, the number infected at each event, follows the logarithmic distribution,

$$\mathbb{P}(Y_i = k) = \frac{-1}{\ln(1-p)} \frac{p^k}{k}, \qquad k \in \{1, 2, 3, \ldots\}. \tag{2}$$

The characteristic function for $Q(t)$ reads,

$$\phi_{Q(t)}(u) = \mathbb{E}[e^{iuQ(t)}] = \exp\left(rt \ln\left(\frac{1-p}{1-pe^{iu}}\right)\right) = \left(\frac{1-p}{1-pe^{iu}}\right)^{rt}, \tag{3}$$

with $\lambda = -r\ln(1-p)$ and thus $Q(t)$ follows a negative binomial process,

$$Q(t) \sim \mathrm{NB}(rt, p). \tag{4}$$

In this process, during a communicable period, t, an infected individual infects $Q(t)$ people based on a draw from the negative binomial with mean $rtp/(1-p)$. The infection events occur continuously in time according to the Poisson arrivals. However, the communicable period, t, is in actuality a random variable, T, which we model as a gamma process[1] with density,

$$f_{T(t)}(x) = \frac{b^{at}}{\Gamma(at)} x^{at-1} e^{-bx}, \tag{5}$$

which has a mean of $\bar{T} = at/b$. By promoting the communicable period to a random variable, the negative binomial process changes into a Lévy process with characteristic function,

$$\mathbb{E}[e^{iuZ(t)}] = \exp(-t\psi(-\eta(u))) = \left(1 - \frac{r}{b}\ln\left(\frac{1-p}{1-pe^{iu}}\right)\right)^{-at}, \tag{6}$$

where $\eta(u)$, the Lévy symbol, and $\psi(s)$, the Laplace exponent, are respectively given by,

$$\mathbb{E}[e^{iuQ(t)}] = \exp(t\,\eta(u)) \tag{7}$$

$$\mathbb{E}[e^{-sT(t)}] = \exp(-t\,\psi(s)), \tag{8}$$

$$\tag{9}$$

and so,

$$\eta(u) = r\ln\left(\frac{1-p}{1-pe^{iu}}\right), \tag{10}$$

$$\psi(s) = a\ln\left(1 + \frac{s}{b}\right). \tag{11}$$

$Z(t)$ is the random number of people infected by a single infected individual over her random communicable period. Without loss of generality, we absorb t into a (or alternatively set $t = 1$, representing a single lifetime) giving a mean communicable period $\bar{T} = a/b$. The gamma process smears out the end of communicable period.

We see that $R_0 = \mathbb{E}[Z(1)] = arp/(b(1-p))$, and thus our process has the same mean as the negative binomial process with a fixed stopping time of $t = a/b$. In fact,

[1] We apply a gamma process subordinator to the negative binomial process.

since $\lambda = -r\ln(1-p)$ we have the simple relationship,

$$R_0 = \left(\frac{a\lambda}{b}\right)\left(\frac{-p}{\ln(1-p)(1-p)}\right) \tag{12}$$

$$= \text{Mean number of infectious events in a lifetime} \times \tag{13}$$

$$\text{Mean number infected at each event.} \tag{14}$$

The variance of the our counting process is over-dispersed relative to the the negative binomial,

$$\text{Var}(Z(1)) = \frac{apr}{b(1-p)^2}\left(1+\frac{rp}{b}\right) \tag{15}$$

$$= \text{Var}(\text{NB}(ar/b, p)) + \frac{ar^2p^2}{b^2(1-p)^2}. \tag{16}$$

The model has four parameters:

- p sets the number of infected people per infectious interaction. The mean number of infected people per infectious event is, $\mu = -\frac{p}{(1-p)\ln(1-p)}$.
- $\lambda = -r\ln(1-p)$ gives the arrival rate of infectious events.
- a, b together set the mean communicable period, $\bar{t} = a/b$, and determine the variance along with the skewness and kurtosis of the gamma distribution, $\Gamma(a, b)$. In the limit $b \to 0$ with a/b finite, the gamma distribution becomes a delta function at the mean time and we recover the negative binomial process evaluated at $t = a/b$.

The characteristic function eq.(6) of the stopped stochastic process allows us to explore the model's analytical properties, which can help decision makers better understand trade-offs in small environments.

2.2 Contact tracing and propagation interruption

The process in eq.(6) represents the spread of the disease from an infected individual without any mitigation strategies. Imagine that we can trace, contact, and isolate infected individuals with a success probability q and with an mean time to isolation of $\bar{m} < \bar{t} = a/b$ after the infectious event. We assume that once isolated, there is no chance for the infected individual to spread the disease any further. We again imagine that the isolation time is gamma distributed but with parameters (a', b') leading to the isolation process, $Z'(\bar{1})$,

$$\mathbb{E}[e^{iuZ'(1)}] = \left(1 - \frac{r}{b'}\ln\left(\frac{1-p}{1-pe^{iu}}\right)\right)^{-a'}. \tag{17}$$

Notice that the branching process for a successful isolation event has the same form as the original process with a mean time of the random communicable period of $\bar{m} = a'/b'$.

Thus, the trace-contact-isolate branching process becomes,

$$N = \prod_{j=0}^{1} [Z_j(1)]^{\mathbb{I}\{y=j\}}; \quad y \sim \text{Bin}(1, q); \tag{18}$$

$$\mathbb{E}(N) = q\,\mathbb{E}(Z(1)) + (1-q)\,\mathbb{E}(Z'(1)), \tag{19}$$

where N is the number of infections produced by an infected person during her communicable period, and q is the probability of a successful isolation event. Instead of arbitrarily cutting the communicable period's density function based on an isolation event as prescribed in [11], our model maintains the same form of the generating function throughout by shifting the mean of the communicable period's gamma process. In a contact-trace-isolate policy, the expected number of infections per infected individual becomes,

$$R_{\text{effective}} = \mathbb{E}[N] = q\,\mathbb{E}[Z'(1)] + (1-q)\,\mathbb{E}[Z(1)], \tag{20}$$

$$= \left(q\left(\frac{a'/b'}{a/b} - 1\right) + 1\right)\frac{(a/b)rp}{(1-p)} \tag{21}$$

$$= \underbrace{\left(q\left(\frac{a'/b'}{a/b} - 1\right) + 1\right)}_{\text{suppresion factor}} \underbrace{\left(\frac{-p}{\ln(1-p)(1-p)}\right)((a/b)\lambda)}_{R_0} \tag{22}$$

$$= \underbrace{\left(q\left(\frac{a'/b'}{a/b} - 1\right) + 1\right)}_{\text{suppresion factor}} \times \tag{23}$$

$$\text{Mean number infected at event} \times \text{Mean number of events.} \tag{24}$$

Eq.(24) provides intuition for comparing competing courses of action by affording trade-off analyses. In a lock-down, health authorities control the spread of the disease by lowering the human-to-human interaction rate λ. If λ can be made sufficiently small, R_0 will drop below unity and the outbreak will come under control. The first term in eq.(24) represents a suppression factor, which by construction is less than unity, and results from an isolation policy with success probability q. Alternatively, that same reduction in R_0 can also be achieved by a lock-down scenario if the infection event rate, λ, is reduced[2] by the same suppression factor. Thus, we see an equivalence in generating $R_{\text{effective}}$ from the two different mitigation strategies, each of which may come at different economic costs.

To make the observation concrete, suppose $\lambda = .20$ implying an average of 0.20 infectious events per day, $p = 0.5$ implying an average of 1.44 infections per infectious event, and a mean communicable period of $a/b = 5.5$ days. The parameters imply $R_0 = 1.59$. Figure 1 shows iso-contours of fixed suppression factor in the $a'/b' - q$ plane. We can now see the trade-off between a lock-down policy with a fixed suppression factor

[2]We recognize that a lock-down would probably reduce p in the logarithmic distribution as well, but we suspect that effect is secondary. We suspect that p, which sets the number of people infected during an event, is not nearly as sensitive to a lock-down scenario as compared to the expected number of events during the communicable period.

and a contact tracing with isolation policy which generates the suppression factor from successful contact tracing events. For a fixed suppression factor figure 1 shows the equivalent curve in the $a'/b' - q$ plane. The economic costs of generating the same value the suppression factor among the two strategies (lock-down vs contact tracing with isolation), with its corresponding reduction in the effective R_0, will in general not be the same. From the figure we see in this example that a contact tracing with isolation policy with an isolation probability of 0.75 and a mean isolation time of 4 days is equivalent to reducing the rate of human-to-human infection events by a factor of approximately 1.25. A lock-down that reduces human interactions by a factor of 1.25 will almost certainly cost much more than the corresponding contact tracing with isolation strategy [9].

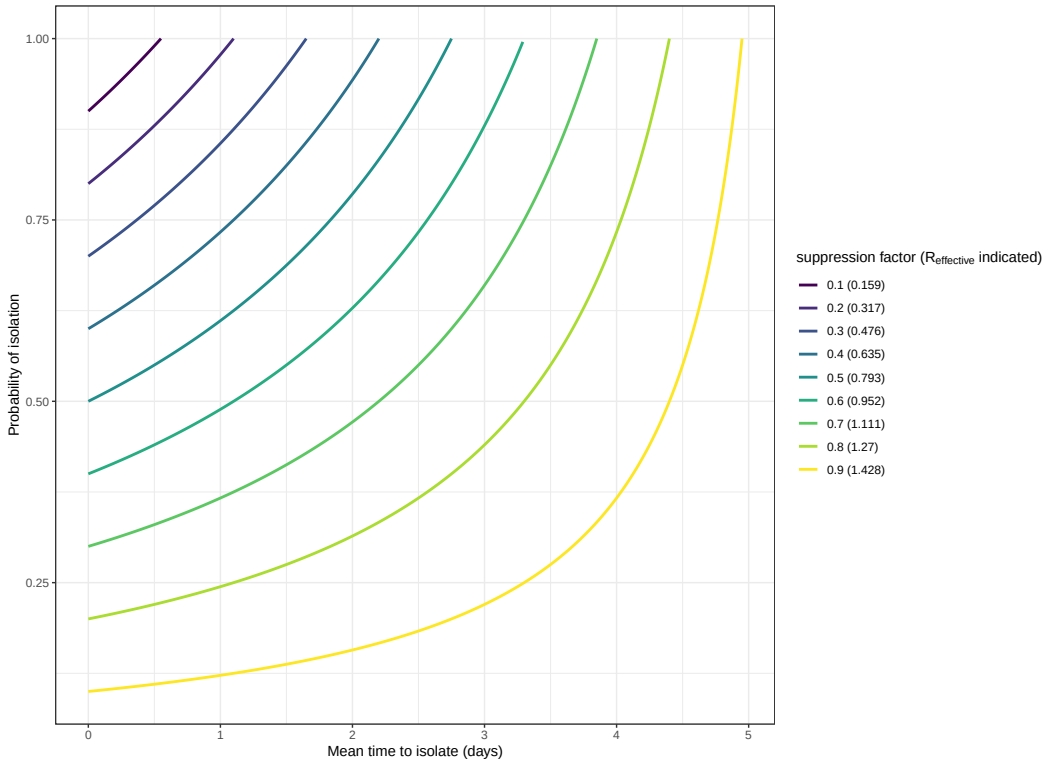


Figure 1: Iso-contours in the $a'/b' - q$ plane with equivalent lock-down factor f.

2.3 Renewal equations and Malthusian parameters

Given a random characteristic $\chi(t)$, such as the number of infectious individuals at time t, (e.g., $\chi(t) = \mathbb{I}(t \in [0, \lambda_x))$ where λ_x is the random communicable period) the expectation of the process follows,

$$\mathbb{E}(Z(t)) = \mathbb{E}(\chi(t)) + \int_0^t \mathbb{E}(Z(t-u))\mathbb{E}(\xi(du)). \tag{25}$$

Defining the Malthusian parameter, $\alpha > 0$, if it exists, by,

$$\int_0^\infty e^{-\alpha t}\mathbb{E}(\xi(dt)) = 1, \tag{26}$$

we can change eq.(25) into a renewal equation,

$$e^{-\alpha t}\mathbb{E}(Z(t)) = e^{-\alpha t}\mathbb{E}(\chi(t)) + \int_0^t e^{-\alpha(t-u)}Z(t-u)e^{-\alpha u}\mathbb{E}(\xi(du)), \tag{27}$$

which has the solution,

$$\lim_{t\to\infty} e^{-\alpha t}\mathbb{E}(Z(t)) = \frac{\int_0^\infty e^{-\alpha u}\mathbb{E}(\chi(u))\,du}{\underbrace{\int_0^\infty ue^{-\alpha u}\mathbb{E}(\xi(du))}_{\beta}}. \tag{28}$$

Thus the asymptotic behaviour of the solution is governed by the pair parameters α, and β.

Recall that $\xi(t) = \#\{k : \sigma(\omega, k) \le t\}$ counts the number of infections during the observation window $[0, t]$ over event space $\Omega(\omega)$. In our model we have,

$$\mathbb{E}(\xi(t)) = \lambda\mu t(1 - G(t)) + \lambda\mu \int_0^t ug(u)\,du, \tag{29}$$

where λ and μ are respectively the Poisson arrival rate and the mean of logarithmic distribution, and where,

$$g(u) = \frac{b^a}{\Gamma(a)}\,u^{a-1}e^{-bu}; \quad G(t) = \int_0^t g(u)\,du. \tag{30}$$

Therefore,

$$d\mathbb{E}(\xi(t)) = \lambda\mu(1 - G(t))\,dt, \tag{31}$$

which leads to the expected result for the mean of direct infections per individual,

$$\int_0^\infty \lambda\mu(1 - G(t))\,dt = \lambda\mu\left(\frac{a}{b}\right). \tag{32}$$

Using eq.(31) and eq.(26) we find that,

$$\alpha = \lambda\mu\left(1 - \left(\frac{b}{\alpha + b}\right)^a\right) \tag{33}$$

$$\beta = \frac{1}{\alpha}\left(1 - \frac{a\lambda\mu}{b}\left(\frac{b}{\alpha + b}\right)^{a+1}\right), \tag{34}$$

which we can solve for the Malthusian parameter, α, by Newton-Raphson. The asymptotic solution to eq.(25) given that the Malthusian parameter exists is,

$$\mathbb{E}(Z(\infty)) \sim \frac{e^{\alpha t}}{\alpha\beta}. \tag{35}$$

If $R_0 < 1$ the branching process will not experience asymptotic exponential growth, instead we can solve eq.(25) for its long term limit,

$$\mathbb{E}(Z(\infty)) = \frac{\mathbb{E}(\chi(\infty))}{1 - \lambda\mu a/b}. \tag{36}$$

2.4 Extinction probabilities and component sizes

In the CMJ framework, we have the generating function,

$$G(t;s) = \mathbb{E}\left[s^{Z(t)} \mid Z(0) = 1\right], \tag{37}$$

with the number of infected by time t composed of infections at time σ_i,

$$\xi(t) = \#\{\sigma_i : \sigma_i \leq t\} = \sum_{i=1}^{\infty} \mathbb{I}\{\sigma_i \leq t\}. \tag{38}$$

The probability of extinction reads,

$$Q = \mathbb{P}\left(\lim_{t\to\infty} Z(t) = 0\right) = \lim_{t\to\infty} G(t;0) \tag{39}$$

$$Q = \lim_{t\to\infty} \mathbb{E}\left[\prod_{i=1}^{\xi(t)} G(t-\delta_i;0)\right] = \mathbb{E}\left[Q^N\right], \tag{40}$$

and for our model, we arrive at the transcendental relationship,

$$Q = \left(1 - \frac{r}{b}\ln\left(\frac{1-p}{1-pQ}\right)\right)^{-a}. \tag{41}$$

Again, we can apply Newton-Raphson and solve for the extinction probability Q.

In addition to the extinction probability for our branching process, we can estimate the average number of total infected people at extinction if extinction occurs by considering the theory random graphs. The branching process is a directed bipartite graph (it is a tree) and given the generating function for the process, we know the distribution of the outgoing edges from a randomly chosen vertex. In [15], the authors extend Erdos-Renyi constructions of random graphs to graphs with arbitrary vertex degree. They compute the mean component size for graphs, including graphs excluding the giant component, if it exists.

The total number infected corresponds to the random characteristic $\mathbb{E}(\chi(t)) = 1$ and thus eq.(36) has the non-Malthusian growth solution,

$$\mathbb{E}(Z(\infty)) = \frac{1}{1 - \lambda\mu a/b}. \tag{42}$$

In [15], the authors consider two generating functions,

- $G_0(s)$: the generating function for the probability distribution of the vertex's degree; and
- $G_1(s) = G_0'(s)/G_0'(1)$: the generating function for the probability distribution of the outgoing edges from a randomly chosen vertex.

Eq.(6) with $e^{iu} \to s$ is $G_1(s)$ in the notation of [15] and for our purposes we do not need an explicit formula for $G_0(s)$. The average component size of the graph, in the absence of a giant component, is [15]

$$\bar{x} = 1 + \frac{G_0(1)'}{1 - G_1'(1)}, \tag{43}$$

which matches the renewal equation solution eq.(42) if $G_0'(1) = G_1'(1)$—that is, the generating two functions intersect tangentially at $s = 1$.

At $G_1'(1) = 1$ a phase transition occurs and the giant component emerges. The fraction of the graph occupied by the giant component is,

$$S = 1 - G_0(Q), \tag{44}$$

where Q is the extinction probability for the distribution of outgoing edges, $Q = G_1(Q)$. Since the fraction of the graph that does not belong to the giant component is composed precisely of those graphs which have gone extinct in our process, we impose $G_0(Q) = Q$. Thus, we demand that the two generating functions intersect on the 45° line at $Q < 1$ when $\alpha > 0$. The average component size in this case becomes [15],

$$\bar{x} = 1 + \frac{zu}{1 - G_1'(1)}, \tag{45}$$

where,

$$z = \frac{1 - Q}{\int_u^1 G_1(s)\, ds}. \tag{46}$$

As $Q \to 1$ we see that $G_1(Q)$ and $G_0(Q)$ increasingly intersect tangentially, finally becoming tangent at $Q = 1$, which is consistent with our observation in eq.(43). We take $\bar{x}$ to be the average size of the total infected population at extinction, if extinction occurs.

3 A scenario planing exercise: Policy input for return to work

One area of application for our model is helping decision makers understand counterfactual outcomes in a return-to-work policy exercises. In setting policies, decision makers must weigh the operational needs of their business while against the possibility of an outbreak in the work environment. In addition to the analytical results that our model provides, simulation can further help ring-fence difficult decisions.

Our model requires four parameters, the arrival rate of infectious interactions, the average number infected at each event, and two parameters which govern the communicable period's density function. We use the open literature [16] as a guide to fix the communicable period; we fix $a = 4.66$ days with $b = 0.85$; these parameter choices give a mean communicable period of 5.5 days with 97.5% of the communicable period ending in 11.5 days. In figure 2, we show the density function arising from our parameter settings. The decision maker has control over the remaining two parameters. By limiting meeting sizes, restricting the number of employees interactions, and by mandating the use of personal protective equipment, the decision maker can set the variables controlling the arrival rate of infectious events and the number infected at each event. We treat the population as homogeneous, holding fixed the arrival rate for infections and the number infected per event over time. In a small setting, in reality, we expect that as people become infected the social network will change, even in the limit of an unmitigated outbreak. Those changes which will have an effect on the basic parameters of our branching model as the population becomes infected, but exactly

how the network changes is a complicated phenomena. Feedback can move the arrival rate and the number infected at each event in competing directions. By ignoring any time dependence in the basic parameters, our model provides a baseline understanding on how COVID-19 propagates in a small populations.

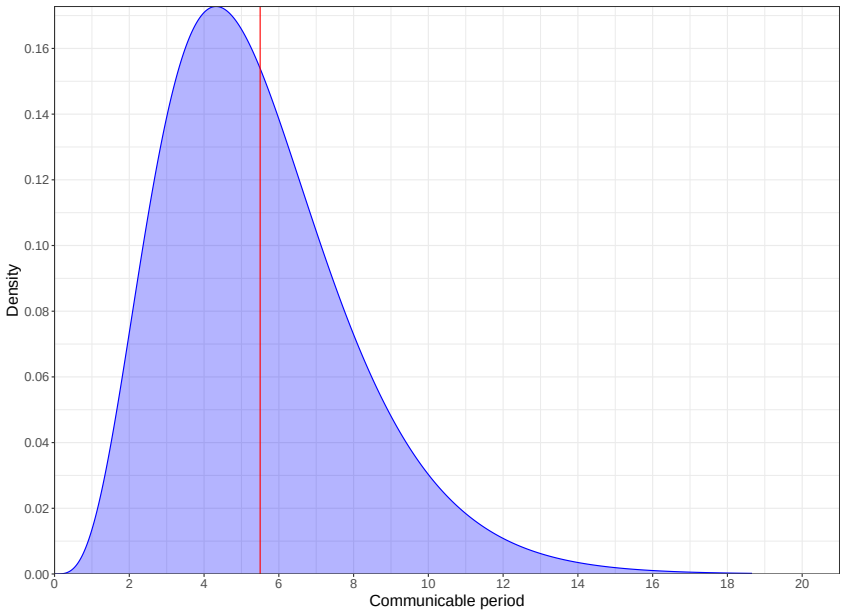
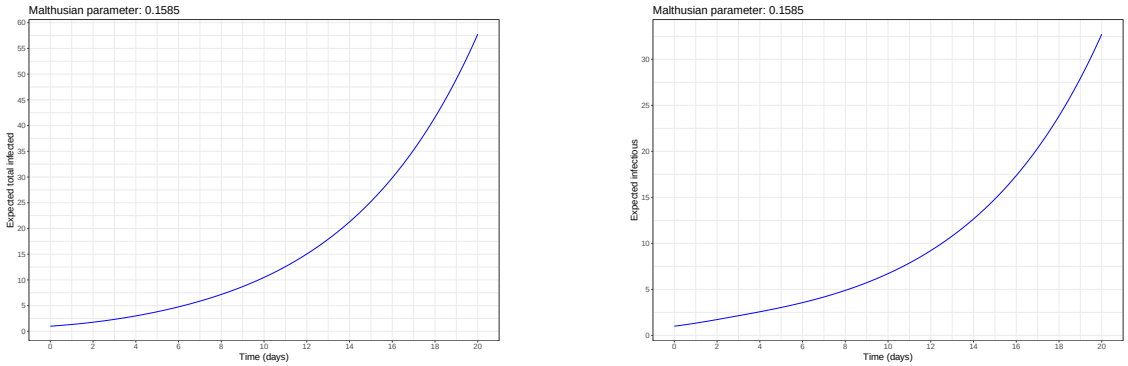


Figure 2: The communicable period: $a = 4.7$, $b = 0.85$. The mean of 5.5 days is indicated by the read vertical line.

Imagine a scenario in which a decision maker has an office space with 100 employees and she must decide on mitigation strategies. Suppose a baseline scenario with $\lambda = 0.2$ and $p = 0.5$ (corresponding to an average of 1.44 infected per event). Given the properties of the communicable period, this scenario corresponds to $R_0 = 1.59$, which implies that if an infected person arrives in the population, in expectation, the branching process will lead to exponential growth in infections. In figures 3a and 3b we display the solution to the renewal equation with this parameter choice for the expected number of infected people and the expected size of the active infectious population respectively.

Let us suppose that the decision maker can change the model parameters λ and p through policy considerations, creating two possible alternative scenarios, each coming at different financial costs. Our model allows the decision maker to investigate trade-offs between starting from one undetected infected individual in the workplace. We summarize model outputs between two scenarios in table 1.

In some office environments, we can imagine a scenario in which management introduces an aggressive testing scheme to isolate infected employees. Suppose our manager faces the baseline scenario on 1 but instead of manipulating interaction rate or meeting sizes, the manager implements a test with a 90% chance of a successful isolation and sharply peaked at a mean of three days. In figure 4 we display the density of the communicable period in the presence of a successful isolation event. Using eq.(24) we see that $R_{\text{effective}} = 0.59 \times 1.59 = 0.93$, and thus this isolation strategy turns an



(a) Total infections (b) Active infections

Figure 3: The expected number of total and active infections as function of time in the baseline planning scenario.

		scenario properties		
description	symbol	baseline	scenario 1	scenario 2
Poisson arrival rate	λ	0.20	0.24	0.1
Logarithmic distribution parameter	p	0.50	0.30	0.5
communicable period shape parameter	a	4.7	4.7	4.7
communicable period rate parameter	b	0.85	0.85	0.85
mean number of new infections per infected individual	R_0	1.59	1.59	0.79
extinction probability	Q	0.61	0.54	1
mean size at extinction	$\bar{x}$	3.1	2.9	4.8
mean number infected at after one week	$N_{1\mathrm{w}}$	6	6	2.3
mean number infected at after two weeks	$N_{2\mathrm{w}}$	22	23	3.3

Table 1: Model properties of the planning scenarios. Each scenario starts with one undetected infected individual.

exponentially growing configuration into a process that will go extinct almost surely. Figure 5 shows 10,000 sample paths of the isolation process over 100 days. Most paths go extinct within two weeks and the average total number of infected is 18 people.

4 A note on parameter inference and an example with US county data

This paper describes a gamma negative binomial branching process (GNBBP) on the number of new infections generated by an infected individual. Given a set of observed $\{n_k\}_{k=1}^{K}$ infection counts for K individuals, a complete Bayesian analysis of the model

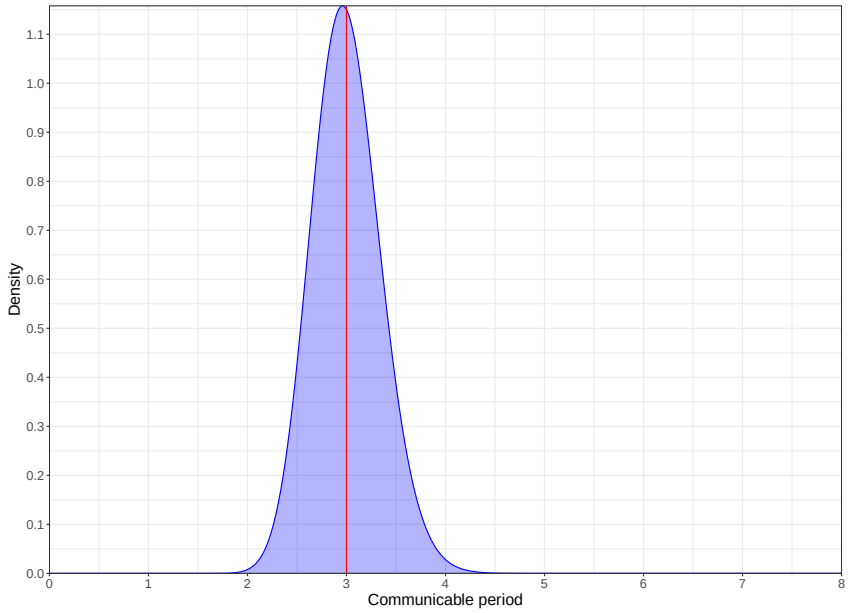


Figure 4: The communicable period with a successful isolation event ($a = 75, b = 25$).

is possible, in which all model parameters are identifiable, using for example, the infrastructure provided in [17]. Under this scheme, all four parameters (r, p, a, b) can be resolved, allowing for a full posterior predictive analysis.

Define a complete history of an outbreak as a set of N observations taking the form of a 6-tuple:

$$(i, j, B_i, D_i, m_i, o_i), \tag{47}$$

where

i	index of individual
j	index of parent
B_i	time of birth
D_i	time of death
m_i	number of offspring birth events
o_i	number of offspring.

With the following summary statistics

$$L = \sum_i D_i - B_i \qquad \Lambda = \prod_i (D_i - B_i) \qquad M = \sum_i m_i \qquad O = \sum_i o_i$$

we can build a Gibbs sampler over the GNBBP parameters as follows:

$$\begin{aligned} p \,|\, r, L, O &\sim \text{Beta}\,(a_0 + O, b_0 + rL) \\ r \,|\, p, L, M &\sim \text{Gamma}\,(\eta_0 + M, \rho_0 - L\log(1-p)) \\ b \,|\, a, L, N &\sim \text{Gamma}\,(\gamma_0 + aN, \delta_0 + L) \\ a \,|\, b, \Lambda, N &\sim \text{GammaShape}\,(\epsilon_0\Lambda, \zeta_0 + N, \theta_0 + N) \end{aligned} \tag{48}$$

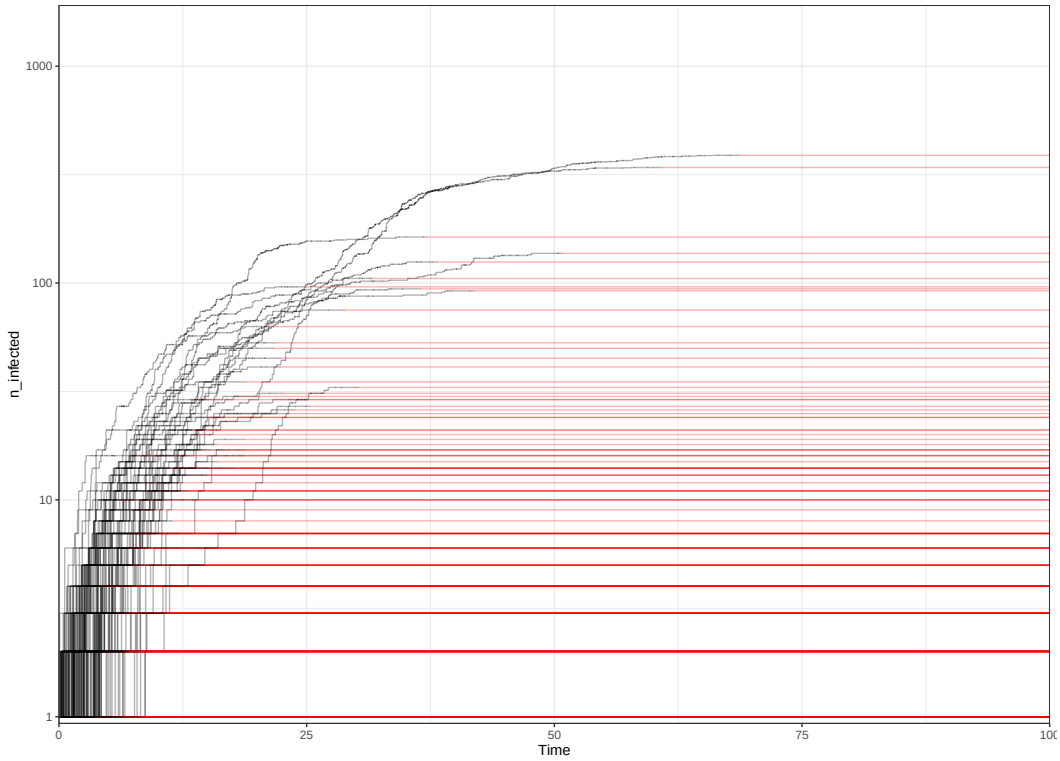


Figure 5: A contract tracing with isolation strategy: 10,000 simulations over 100 days. The probability of successful isolation is $q = 0.90$. Red lines indicated extinct paths from the moment of extinction. All paths eventually go extinct as the result of the intervention.

where $a_0, b_0, \eta_0, \rho_0, \gamma_0, \zeta_0, \epsilon_0, \theta_0$ are hyper-parameters.

Unfortunately, under real world conditions, we are rarely fortunate enough to have such complete and pristine data. Readily available COVID-19 data almost always takes the form of cumulative case count data by geographic region, but if public health officials can collect data in the form of eq.(47) during a local outbreak, Gibbs sampling will yield posteriors for all model parameters. The underlying propagation mechanism of the GNBBP affords additional interpretability to the model, which, in turn, facilitates incorporation of other prior information. For example, knowing that it is unlikely that multiple thousands of individuals could be infected in a single interaction allows us to set a prior with more mass on values of p closer to 0; moreover, direct experimental measurement of this parameter might be possible in a lab setting or augmented by fine grained clinical data. Similar considerations apply to the rate of infectious events, λ. The parameters which govern the communicable period, (a, b), can be inferred from clinical observations. Likewise, information on probable ranges of R_0 from other comparable infections could also be leveraged to provide a joint constraint on r, p, a, and b.

Even with limited data we can still estimate parts of the model. In particular, we can estimate the Malthusian parameter of eq.(26) from cumulative count data that exhibits exponential growth by applying the asymptotic solution, eq.(35). Since the

Malthusian parameter depends on the product of the infection arrival arrival rate an the average number infected per event, $\lambda\mu$, an estimate of the Malthusian parameter yields an estimate of R_{eff} through the parameters a and b of the communicable period's gamma distribution,

$$\begin{aligned} R_{\text{eff}} &= \lambda\mu \cdot \frac{a}{b}, \\ &= \frac{a\alpha}{b\left[1 - \left(\frac{b}{\alpha+b}\right)^a\right]}. \end{aligned} \tag{49}$$

Based on the clinical literature [16], we take $a = 4.66$ and $b = 0.85$ giving a mean communicable period of approximately 5.5 days with a 97.5% of 11.5 days.

The New York Times provides a COVID-19 case count dataset for the United States resolved on the county level [18]. A team at the New York Times curates the data from multiple sources and ensures data accuracy. Using the New York Times data, we estimate the Malthusian parameter for US counties which exhibit exponential growth over the period July 1, 2020 to August 20, 2020. We use a hierarchical Bayesian construction with a county level random effect,

$$\begin{aligned} \log(Z) &= \alpha t + \gamma + a_i t + g_i + \epsilon \\ a_i &\sim \mathrm{N}(0, \sigma_1^2) \\ g_i &\sim \mathrm{N}(0, \sigma_2^2) \\ \epsilon &\sim \mathrm{N}(0, \sigma^2), \end{aligned} \tag{50}$$

where i is the county label; the variance parameters use half-Cauchy priors and the fixed and random effects use normal priors. We estimate the model and generate posterior distributions for all parameters using JAGS [19]. The posterior means of the Malthusian parameter for each county gives R_{eff} over the time interval through eq.(49). We display the US county results for R_{eff} in figure 6. Over the mid-summer, we see that the geographical distribution of R_{eff} across the US singles out the Midwestern states and Hawaii as hot-spots while Arizona sees no county with exponential growth.

5 Discussion

In small setting with localized outbreaks, a branching model offers a stochastic view of the propagation. To be useful in a decision making setting, the branching model must be parsimonious yet contain appropriate features which match clinical observations and bounds on key parameter such as R_0.

Our model contains physically motivated mechanisms that link to macroscopic observables. For instance, our model generates the negative binomial count process by coupling Poisson infectious event arrivals with the logarithmic for the number infected at each event. We extend the model of [11] by including the serial interval distribution within a complete generative continuous time stochastic branching process. Furthermore, our model allows for an exploration of trade-offs between mitigation strategies at the microscopic level, especially in light of the model's analytical tractability. Because our

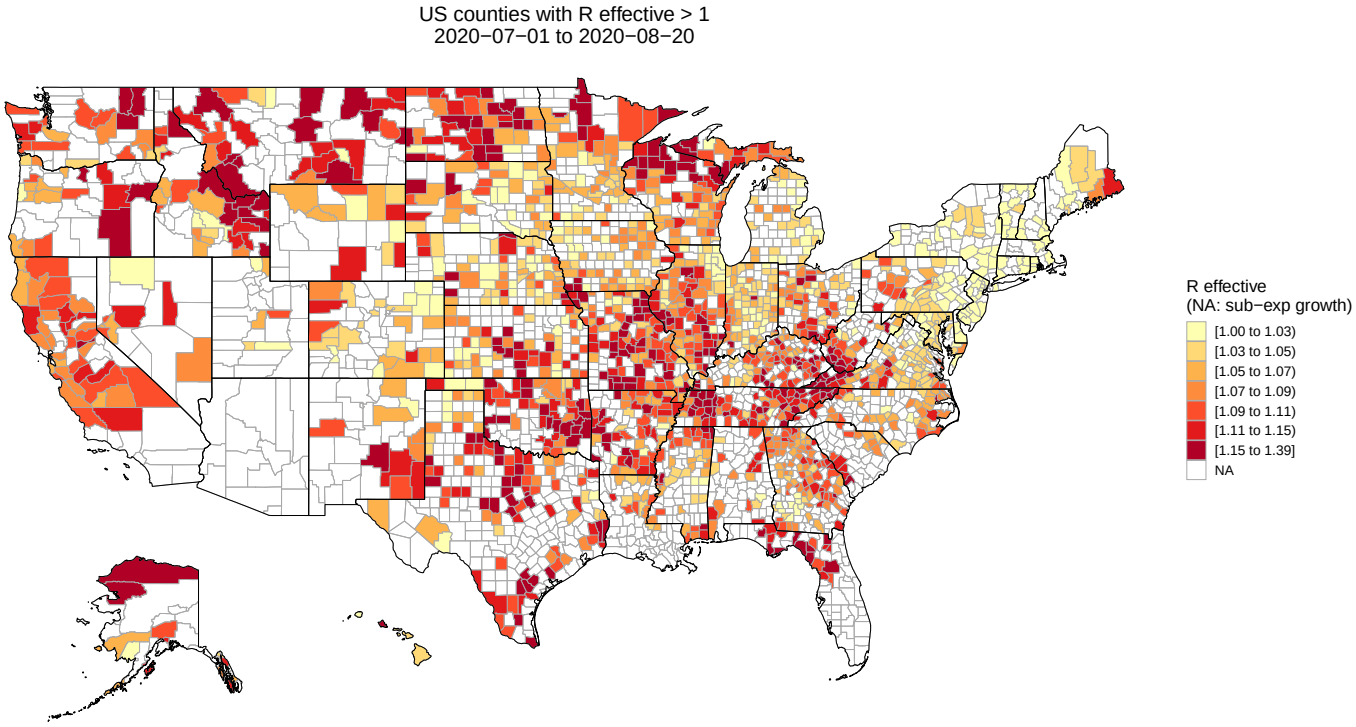


Figure 6: Summer 2020 geographical distribution of R_{eff} across the United States: 2020-07-01 to 2020-08-20.

model includes the generating function of the underlying branching process, it easy to build a continuous time simulation engine, model the effect of intervention strategies, and estimate model parameters through Bayesian hierarchical methods.

References

[1] Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whittaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick GT Walker, Han Fu, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, 2020.

[2] Marc Lipsitch. Estimating case fatality rates of covid-19. *The Lancet Infectious Diseases*, 2020.

[3] John P. A. Ioannidis, Cathrine Axfors, and Despina G. Contopoulos-Ioannidis. Population-level covid-19 mortality risk for non-elderly individuals overall and for non-elderly individuals without underlying diseases in pandemic epicenters. *medRxiv*, 2020.

[4] Eran Bendavid, Bianca Mulaney, Neeraj Sood, Soleil Shah, Emilia Ling, Rebecca Bromley-Dulfano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra, James Tedrow, Dona Tversky, Andrew Bogan, Thomas Kupiec, Daniel Eichner, Ribhav Gupta, John Ioannidis, and Jay Bhattacharya. Covid-19 antibody seroprevalence in santa clara county, california. *medRxiv*, 2020.

[5] Tianbing Wang, Zhe Du, Fengxue Zhu, Zhaolong Cao, Youzhong An, Yan Gao, and Baoguo Jiang. Comorbidities and multi-organ injuries in the treatment of covid-19. *The Lancet*, 395(10228):e52, 2020.

[6] Covid-19 pandemic planning scenarios. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html. Accessed: 2020-06-20.

[7] John Ioannidis. The infection fatality rate of covid-19 inferred from seroprevalence data. *medRxiv*, 2020.

[8] The Lancet. Sustaining containment of covid-19 in China. *Lancet (London, England)*, 395(10232):1230, 2020.

[9] Daron Acemoglu, Victor Chernozhukov, Ivn Werning, and Michael D Whinston. Optimal targeted lockdowns in a multi-group sir model. Working Paper 27102, National Bureau of Economic Research, May 2020.

[10] Linda JS Allen. A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling*, 2(2):128–142, 2017.

[11] Joel Hellewell, Sam Abbott, Amy Gimma, Nikos I Bosse, Christopher I Jarvis, Timothy W Russell, James D Munday, Adam J Kucharski, W John Edmunds, Fiona Sun, et al. Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 2020.

[12] Ronald A Fisher, A Steven Corbet, and Carrington B Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.

[13] Maurice H Quenouille. A relation between the logarithmic, poisson, and negative binomial series. *Biometrics*, 5(2):162–164, 1949.

[14] David Applebaum. *Lévy processes and stochastic calculus*. Cambridge university press, 2009.

[15] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2), Jul 2001.

[16] Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582, 2020.

[17] Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2013.

[18] The New York Times. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html, 2020.

[19] Martyn Plummer. Jags: A program for analysis of bayesian graphical models using gibbs sampling, 2003.