Epidemics 44 (2023) 100708

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

The effective reproductive number: Modeling and prediction with application to the multi-wave Covid-19 pandemic

Razvan G. Romanescu [a,b,*], Songdi Hu [c], Douglas Nanton [b], Mahmoud Torabi [a], Olivier Tremblay-Savard [c], Md Ashiqul Haque [a]

[a] Department of Community Health Sciences, University of Manitoba, Canada
[b] Center for Healthcare Innovation, University of Manitoba, Canada
[c] Department of Computer Science, University of Manitoba, Canada

ARTICLE INFO

Keywords:
Effective reproductive number
Infectious disease transmission
Population network
Covid-19
SIRS compartmental models

ABSTRACT

Classical compartmental models of infectious disease assume that spread occurs through a homogeneous population. This produces poor fits to real data, because individuals vary in their number of epidemiologically-relevant contacts, and hence in their ability to transmit disease. In particular, network theory suggests that super-spreading events tend to happen more often at the beginning of an epidemic, which is inconsistent with the homogeneity assumption. In this paper we argue that a flexible decay shape for the effective reproductive number (R_t) indexed by the susceptible fraction (S_t) is a theory-informed modeling choice, which better captures the progression of disease incidence over human populations. This, in turn, produces better retrospective fits, as well as more accurate prospective predictions of observed epidemic curves. We extend this framework to fit multi-wave epidemics, and to accommodate public health restrictions on mobility. We demonstrate the performance of this model by doing a prediction study over two years of the SARS-CoV2 pandemic.

1. Background

Since the introduction of the susceptible-infected-removed (SIR) model of epidemic spread nearly a century ago (Kermack and McKendrick, 1927), the homogeneity of individuals in terms of their chance of interacting with one another has been a mainstay assumption in infectious disease modeling to this day. This is despite ample evidence that some individuals contribute much more to disease transmission than others (Lau et al., 2017; Wang et al., 2020; Petersen et al., 2020), and that the assumption of homogeneous mixing, also called mass action, produces poor fits to real data (Stack et al., 2013; Romanescu and Deardon, 2017). We identify two reasons why the mass action assumption continues to be used in practice. First, a homogeneous population leads to simple mathematical ordinary differential equation (ODE) models that have a physical analog in the time progression of chemical reactions (where the mass action principle also originates). Second is the fact that a heterogeneous model of spread requires additional assumptions about how individuals differ in their ability to transmit disease. This extra information may be either unavailable, or, when present, can increase the complexity and computational burden of the models far beyond simple ODEs (such as, for instance, agent-based models e.g., Deardon et al., 2010).

Network models are one alternative to mass action models. They describe the population via a network with individuals as nodes and contacts as edges, and model individual heterogeneity via a different number of infectious contacts for each individual. A network with first order properties is characterized by the distribution of a node's degree (the number of edges emanating from it). Higher order properties relate to community structures beyond each node's own contacts. We will restrict our attention to first order properties, as this provides a rich enough family of models to work with. First analyzed mathematically in Newman (2003), epidemics over networks have some unique properties, for instance an epidemic can occur for arbitrarily low transmissibility, provided that the degree distribution is skewed enough (heavy tailed). In practice, network models are adept at explaining patterns in real data that cannot be accounted for by mass action models, such as sharp peaks of the epidemic curve, and an initial growth rate that exceeds that predicted using textbook values of the basic reproductive number (R_0) for a disease.

While network models offer much flexibility to explain observed

* Corresponding author at: Department of Community Health Sciences, University of Manitoba, Canada.
E-mail address: Razvan.Romanescu@umanitoba.ca (R.G. Romanescu).

https://doi.org/10.1016/j.epidem.2023.100708
Received 23 December 2022; Received in revised form 4 July 2023; Accepted 13 July 2023
Available online 20 July 2023

data, they have the obvious disadvantage of requiring knowledge of the underlying network, at least as the probability distribution of contacts per individual. This is difficult to observe, and although it has been attempted in limited contexts (staff in a healthcare setting, see e.g., Hornbeck et al., 2012; Machens et al., 2013), these do not easily generalize to the entire population. Fortunately, the flexibility of network models (defined by the first degree, without higher order structure) can be captured via a non-linear decay of the effective reproductive number (R_t), defined as the number of new infections arising from one infected individual at time t. Thus, focusing modeling efforts on estimating R_t will unlock the benefits of network models without the data requirements of defining a network, which is sensible if the end goal is modeling or prediction of case counts. While the idea that a flexible form for R_t – either explicit or implied – can improve realism in an SIR model has been considered before in various ways, this paper takes a unique approach of relating the contact rate between infections and susceptible individuals to the total susceptible fraction remaining in the population (S_t). In particular, we argue that time dynamics of R_t depend on this fraction alone. This idea arises not only from the theory of first-order network models, but also from ODE models of heterogeneous susceptibility. We show how this formulation can be used to model multiple waves of a pandemic, as well as non-pharmaceutical public health interventions, such as lockdowns or capacity limits. Modeling efforts to address these phenomena in the literature tend to rely exclusively on a homogeneous understanding of the population. Having a theory-informed framework for heterogeneity provides a more solid foundation for understanding these complex processes, and leads to improved prediction performance.

The paper is organized as follows: in Section 2 we present the method, including fitting to count data. In Section 3 we fit the model to COVID-19 infection counts from the state of New York and the city of Winnipeg and infer the R_t curve for that population using a few proposed shapes. We also run a prediction study to demonstrate the utility of the R_t curves in predicting the future epidemic trajectory. We conclude in Section 4 by summarizing the benefits of this approach, and discussing the limitations and potential directions for future research.

2. Methods

2.1. Existing approaches to model heterogeneity

Different classes of models have been proposed in the literature to relax the mass action assumption. Compartmental models formulated in an ODE framework are perhaps the most widely-used infectious disease model class, and it is relatively easy to adapt these models to imply a different decay pattern in R_t than under homogeneity. In the equation for new infections, $dI_t/dt = aI_tS_t$, where the product I_tS_t is reflective of the mass action assumption, some have proposed other forms, a popular choice being $dI_t/dt = aS_t^pI_t^q$ (Roy and Pascual, 2006; Stroud et al., 2006; Hethcote and van den Driessche, 1991). Others (Chowell et al., 2016) have identified an exponential decay shape, the most general form of which is $dI_t/dt = aI_tS_t\exp(-bI_t^n)$, as in Granich et al. (2009). It is worth noting that the last equation has been proposed to model HIV transmission in the context of an SI (as opposed to an SIR) population. Sexually transmitted diseases are known to be strongly heterogeneous in terms of contact patterns between individuals (Liljeros et al., 2001). Although these models are not mechanistic, it has been shown that certain mechanistic models can be reduced to a homogeneous model with a nonlinear transmission function. For instance, a randomly mixing population of heterogeneous individuals having distributed infectivity/susceptibility (Novozhilov, 2008); and network models of first degree (Volz, 2008; Miller et al., 2012), where individuals do not mix randomly, but preferentially, according to their degree.

A second approach, which we might call structural, attempts to model heterogeneity explicitly. In an effort to describe the spread of gonnorhea, Hethcote and Yorke, (2014) introduced two models. One model had a core group, which would be responsible for most of the transmissions. While this is in line with observations that a small fraction of individuals are responsible for a disproportionately large number of infections, one core group is likely not accurate enough to account for heterogeneity within this group. Their second model contains N groups, where individuals within each group have similar contact patterns. This model is described by N differential equations, which, while general enough, is probably cumbersome to deal with in practice. Network models, which explicitly describe the contact patterns between individuals, are consistent with the N group approach of Hethcote & Yorke if we regard a group as individuals having the same degree, but have reduced the dimensionality of the problem from N (which could be very high) to only a couple of equations via the use of generating functions (Newman, 2002; Volz, 2008; Miller, 2011). The major weakness of network models, besides being generally more involved mathematically, is that they require a specification of the entire network architecture. First-order network models, to which the paper will refer exclusively henceforth, only require the degree distribution, however even this is not observable, in general. Estimating individual contacts has traditionally been a difficult problem and, although a full review of the topic is outside the scope of this paper, we will only mention two sources of error. One stems from the difficulty in defining what a potentially infectious contact means. Previous studies have attempted to infer contacts either directly, from surveys about the frequency of conversations or physical contacts involving skin-to-skin contact (Bansal et al., 2010; Mossong et al., 2008; Read et al., 2008; Beutels et al., 2006; Wallinga et al., 2006); or indirectly, from cell phone data, i.e., the time that a device is physically away from the user's residence (Chang et al., 2021). Another source of errors can come from the administration of surveys; Feehan & Mahmoud (2021) point to the possibility of recall bias when participants are asked about contacts from the past day(s), as well as social desirability bias, e.g., participants may avoid disclosing behaviors in violation of social distancing policies. Thus, the reliance of network models on a correctly specified degree distribution is a hurdle, and is possibly why their impact has been mostly theoretical.

To conclude this review, on the spectrum of structural vs. phenomenological models, there are the extremes, which we only mention but not consider further. At one end of the spectrum, there are purely phenomenological models, such as the generalized Richard's model (Turner et al., 1976), which seek to fit the cumulative infection counts C_t without specifying dynamics in terms of I_t and S_t, i.e., these models are content to fit observed data without explaining it. At the other end of the spectrum, there are the agent based models, which are data-intensive and can offer a very granular view of each individual and their environment, down to the building level. With enough detail, this class of models will obviously account for heterogeneity, however, besides the difficulty in obtaining epidemiologically-relevant contact data, the computational demands of these models put them out of reach of most users.

2.2. Prediction model

We formulate our prediction model in an SIR framework discretized at an appropriate time step (usually one day). We start from an estimator of the effective reproduction number (Nishiura and Chowell, 2009; Cori et al., 2013): $\widehat{R_t} = \frac{y_t}{\sum_{s=1}^{M} y_{t-s}w_s}$, where y_t is the number of infections on day t; w_s is the generation interval, which gives the probability that the time between a primary and secondary infection is s days; and M is the maximum range of the generation interval. Then we can re-express this formula to predict incidence at time $t+1$ as in Nishiura and Chowell (2009) and Abbott et al. (2020):

$$y_{t+1} = R_{t+1}\sum_{s=1}^{M} w_s y_{t+1-s} \tag{1}$$

The susceptible fraction is updated as in Romanescu and Deardon (2017):

$$S_{t+1} = S_t - \frac{y_{t+1}}{\rho_t N}, \tag{2}$$

where N is the total population size and ρ_t is the underreporting rate, defined as the fraction of reported to actual cases, which is estimated externally. Next, we turn our attention to predicting R_{t+1} in (1). Based on previous work on epidemics in heterogeneous populations, we model the effective reproductive number as

$$R_{t+1} = \alpha S_t c(S_t) \tag{3}$$

where α is the person to person probability of transmission and $c(S_t)$ is an effective contact rate, which decays as S_t decreases (see the Appendix for a review of the theory). By contact we mean all contacts that have the potential to transmit disease, and this is independent of infection status. Intuitively, the reason why is increasing is that more highly connected individuals are more at risk of getting infected compared to those less well-connected. As a result, the average contact rate of infected individuals drops as the epidemic progresses, as does the contact rate of the susceptible set. Thus, people who get infected are both less connected on average than the early infectious individuals and more-connected on average than those who remain susceptible.

Crucially in formulation (3), is a function of S_t alone, and this emerges robustly from the compartmental ODE modeling literature on heterogeneous susceptibility (Novozhilov, 2008). This is also implied by first-order network models built according to the Configuration Model (Molloy & Reed, 1998; Romanescu & Deardon, 2017), thought this might not be robust to higher order network structures. It is worth noting that first-order network models defined by degree distribution alone are more similar conceptually to multi-group compartmental models (Hethcote, Yorke, 2014; Huang et al., 1992; Van den Driessche & Watmough, 2002; etc.) than to general network models. With the latter they only share the degree distribution, but not any higher order properties (such as assortativity, clustering, global network structure, neighborhood effects) or stochasticity. Finally, and from a practical perspective, both key references mentioned that imply Eq. (3) (Novozhilov (2008) and Romanescu & Deardon (2017)) are equivalent to simple ODE models with non-linear transmission rates, in the SIR case.

Functional forms for in Table 1 were chosen to allow for significant flexibility in the shape of R_t, and to accommodate dynamics observed for usual contact distributions as used in network theory. The dynamics of R_t at the start of an outbreak are as follows: initially, when the numbers of infected in the population are low, the epidemic is in a stochastic phase, when there is a non-zero chance of the outbreak dying out. Following this, infection becomes established in the community and spread dynamics begin to follow deterministic trends closely. R_t reaches its highest average value given by formula[1] $R_{\max} = \alpha E[K(K-1)]/E[K]$ (Stack et al., 2012) both in the stochastic phase, and at the start of the deterministic regime, as transmission is driven by superspreaders (i.e., the most connected individuals). After this, R_t starts decaying at a rate that depends on the structure of the network (Newman, 2003; Volz, 2008; Miller et al., 2012). We include two shapes for $\alpha \times c(S_t)$ implied from known degree distributions, namely the Poisson and Geometric (see the Appendix for the exact definitions). The (new) shapes we consider – the two piece Exponential and shifted inverse – allow for potentially high values of R_t at the start of the epidemic, and a varying degree of curvature during the decline. The two piece exponential is meant to accommodate a precipitous decay in R_t at rate a_3 when the population is fully susceptible ($S_t = 100\%$), for the first few percentage points drop in S_t until level $a_2 < 1$; after which the decay proceeds at the much smaller rate a_1. The shifted inverse is another parsimonious solution for producing both a high initial value as well as a kink in the shape of R_t. Since α is not estimable on its own from time series data, we parameterize $\alpha c(S_t)$ as a whole, with parameters that will be estimated.

Table 1
Functional forms of the effective contact rate (times α).

Shape of decay	Parameters	$\alpha \times c(S_t)$
Poisson	$0 \le a_0 \le 1$, a_1(mean of Poisson)	$a_0\{a_1 + \ln(S_t)\}$
Geometric	$0 \le a_0 \le 1$, a_1(parameter of Geometric is $p = 1 - e^{-1/a_1}$)	$2a_0\frac{S_t - 1 + e^{-1/a_1}}{1 - e^{-1/a_1}}$
Two piece exponential	$a_1, a_2, a_3 \quad (a_3 > a_1)$	$e^{a_1 S_t} + \max\left(e^{a_3(S_t - a_2)}, 1\right) - 2$
Shifted inverse	a_1, a_2 $(a_2 > 1)$	$\frac{a_1}{a_2 - S_t}$
Power function	a_1, a_2	$a_1 S_t^{a_2}$
Granich HIV	a_1, a_2, a_3	$a_1 e^{-a_2(1-S_t)^{a_3}}$

We finally include two common shapes of R_t implied by some of the compartmental models reviewed before, namely the power function with $q = 1$ (Stroud et al., 2006; Connell et al., 2009), and the model of Granich et al. (2009). We should explain here how we infer the implied R_t from ODE models. R_t can be identified with the ratio of incidence to prevalence, scaled by the relative infectiousness at time t of the infected individuals (Hens et al., 2012; Thompson et al., 2019; Gostic et al., 2020; Vanni et al., 2021). In a traditional ODE model (Smith and Moore, 2004), this can be written more compactly as $R_t = -(dS_t/dt)/I_t \times D$, where D is the mean duration of infectiousness (according to Gostic et al., 2020), which is constant. Ignoring recoveries, $-dS_t/dt = dI_t/dt$ and thus R_t is proportional to $(dI_t/dt)/I_t$. For example, in the Granich model $R_t \propto S_t \exp(-bI_t^n)$, and, because $S_t + I_t = 1$ in this model, we can write $R_t \propto S_t \exp(-b(1-S_t)^n)$. For consistency with our modeling framework and derivations in the next subsections, we only consider functions of S_t in the shapes.

Together, Eqs. (1 – 3) and the shape of $c(\bullet)$ define the epidemic model used to describe incidence within waves. We will refer to this model as a synthetic network model (SNM), as it is meant to reproduce the behavior of a network model, without explicitly modeling the network. We should clarify that the new shapes we propose are not based on known degree distributions. Our primary purpose here is to find curve shapes that fit epidemiological data well. While we do not make a point of maintaining full consistency with network models, we recognize that doing so is desirable at least on a theoretical level, as this ensures that SNMs continue to have a mechanistic basis. In other words, it would mean that a network can be constructed whose epidemic dynamics are approximated by that SNM. To this end, we have developed a way to derive the degree distribution (if it exists) from an arbitrary curve $c(S_t)$, and this is given in the Appendix.

2.3. Wave definitions

Focusing our discussion on respiratory pathogens, we define the start of a wave to be the day of an observed maximum $\widehat{R_t}$; as mentioned above, this signals that an outbreak has started. The end of a wave will be the day of a minimum in $\widehat{R_t}$, which should be well below one. We do not attempt to make predictions in beween waves. When identifying waves retrospectively, it is important to check visually that the observed maximum $\widehat{R_t}$ does correspond to the start of a downward trend in R_t and is not simply an isolated outlier. Prospectively, we require a way to confirm a recent maximum or minimum. For the maximum we recommend waiting one day to confirm a high observed $\widehat{R_t}$ value as the

[1] Note that in the network modeling literature this is usually given as the formula for R_0. However, strictly speaking this is incorrect. R_0 assumes that there are no other infectious individuals in the population aside from the index case, whereas the formula presented assumes a small number infected of each degree. Thus, we refer to the latter as $R_{\max}$ whereas R_0 is still $R_0 = \alpha E[K]$, where K is an individual's degree.

tentative start of the wave. By 'high' we generally mean a value of around 2 or higher. To limit the chances of a false positive signal for the maximum, we also recommend ensuring that incidence is above a certain threshold (see our previous work in Romanescu and Deardon, 2019, for more details on this approach). For the minimum, we recommend waiting at least 2 weeks for confirmation. This asymmetry reflects the importance of timeliness in predictions at the beginning of a wave, while at the end, the interest is more in ensuring that the wave has ended. Henceforth, we will use retrospectively defined waves, which is fair because here we are more concerned with the times we should start fitting the model, rather than pinpointing the exact start time of a new wave. In practice, the start of a new wave receives much attention and expertise from the broader epidemiological community; thus we make no claims that this definition of waves is the best, except perhaps to say that it makes the most sense to make predictions inside that interval, at least from a network modeling perspective.

2.4. Including restrictions on the social network

In order to accurately fit count data over any length of time beyond the short term (e.g., 2 weeks), it is necessary to accommodate changing restriction measures and other public health interventions, as well as changing transmissibility due to mutations in the pathogen, vaccinations, etc. Our model can accommodate two types of regime shifts: those that affect the underlying network structure, such as workplace closures and limits on gatherings, and those that affect the person-to-person (p2p) probability of transmission α. The latter category is fairly broad, and captures mask mandates, vaccinations, as well as any changes in the pathogen itself, and often combinations of these factors. Changes in p2p probability of transmission result in scaling R_t directly by the amount of relative change, and can be accommodated by including a scale parameter, which multiplies R_t in each wave beyond the first.

Regime shifts that affect the network structure are currently modeled in the literature as a scaling in transmissibility (e.g., Anderson et al., 2020, and others). This is equivalent to keeping the same number of contacts, but reducing the activity of each. An alternative is to keep the strength of contacts the same, but reduce their number, and we choose this approach in this paper. This is arguably more realistic during lockdowns or when modeling restrictions on gatherings, because it effectively reduce the fraction of individuals with many connections, while increasing the fraction with few connections. The limitation of this alternative approach is that every network link in the original network is exactly the same strength. We will model this uneven shift by imposing an exponential tail to the degree distribution of the underlying network. This is equivalent to changing the effective contact rate $c(S_t)$ in the expression of R_t to $c(\psi S_t)$, for an appropriately chosen parameter $\psi < 1$ (see Appendix for a proof of this). The intuitive explanation is that an epidemic that starts on a restricted network has similar dynamics in $c(S_t)$ as one that has progressed on an unrestricted network up to S_{t_0} at time t_0 and we are restarting the count of susceptibles $\widetilde{S}_t$ from 1 at time t_0 by considering the remaining susceptibles as the entire population. The effective contact rate will now be $c(S_{t_0}\widetilde{S}_t)$, for $t > t_0$ and the scaling factor ψ in this case is S_{t_0}. Essentially, restrictions will result in a lower contact rate of the infectious set, one normally associated with a later stage in the epidemic. The expression of R_t under such restriction becomes:

$$R_{t+1} = \alpha S_t c(\psi S_t). \tag{4}$$

One may model less or more severe network restrictions using different ψ values at different times, with a lower value indicating more deleted connections. Parameters $\psi_1 > \psi_2 > \psi_3 > \dots$ corresponding to increasing severity levels $1, 2, 3, \dots$ are included in the estimation scheme described next. We assume a two week adjustment period following the introduction of a new restriction, during which ψ transitions linearly from the old to the new value.

2.5. Estimation procedure

First, let Z_t be the count of all infectors that can cause new infections at time $t+1$ (i.e., counted in y_{t+1}). We can write $Z_t \sim \text{Poisson}(\Lambda_t)$, where $\Lambda_t = \sum_{s=1}^{M} w_s y_{t-s+1}$ is the denominator in the definition of R_t from Section 2.1. Following Romanescu and Deardon (2017), we can write the number of new infections as a sum of secondary infections from each infected individual counted in Z_t, thus: $y_{t+1}|Z_t = \sum_{i=1}^{Z_t} X_{t,i}$, where $X_{t,i}$ are iid expansion factors, or counts of secondary infections for each infected individual. Based on this, the conditional mean and variance given Λ_t are $E(y_{t+1}|\Lambda_t) = E[E(y_{t+1}|Z_t)|\Lambda_t] = \Lambda_t E(X_t)$ and $\text{Var}(y_{t+1}|\Lambda_t) = \Lambda_t\{\text{Var}(X_t) + [E(X_t)]^2\}$, using the conditional version of the law of total variance (Bowsher & Swain, 2012). From previous work, $E(X_t) = R_t$ and a flexible formulation for the variance of X is: $\text{Var}(X_t) = E(X_t)[u + vE(X_t)]$, where u, v will be estimated from data (see Appendix for details). Thus, the conditional count y_{t+1} works out to be:

$$y_{t+1}|\Lambda_t, R_t \sim N(\Lambda_t R_t, \Lambda_t R_t[u + (v+1)R_t]), \tag{5}$$

where normality is justified due to summing over a relatively large number of terms (Λ_t), and also because $X_t < N$ has finite moments. We can now proceed to fit the model parameters using all available case data via maximum likelihood estimation (MLE). To find the MLE, one needs to maximize the log-likelihood (given in the Appendix) over the parameter set, using all the data available at the present time. This is done using the "BFGS" routine of the optim function in R. Standard errors of the estimated parameters are computed using the large sample approximation to the MLE. Namely, the observed Fisher information matrix is obtained as the Hessian of the negative log-likelihood at the MLE, which is returned by the optimizer. This matrix is inverted to yield the variance-covariance matrix of the vector of estimates.

The underreporting rate ρ_t used to update the susceptible fraction in (2) is estimated externally. The problem of estimating underreporting rates is generally difficult, and existing methods (e.g., Wu et al., 2020; Irons and Raftery, 2021) often require strong assumptions that change over time, as public health attitudes towards testing and reporting shift over the course of a pandemic. Although this discussion is not the focus of this paper, it seems to us that anchoring case reports to death counts is more theoretically appealing than using testing rates, which appear to be driven more by politics and changing public attitudes toward enforcement.

2.6. Extension to SIRS: accounting for reinfection

Pathogens such as SARS-CoV2, as well as the seasonal influenza virus (Webster et al., 1992), persist in the population as multiple strains that interact with one another within hosts via cross-immunity, as well as evolve genetically. While most ODE models regard the mechanism of reinfection to be a time decay of immunity – usually exponential – at the host level (e.g., Bjørnstad et al., 2020), this may not be realistic in the case of a multi-strain, evolving pathogen. Here we follow the state space models of Gog & Swinton (2002) and Gog & Grenfell (2002), which are more suitable when reinfections are driven by changes in the pathogen. In this framework (referred to as polarized immunity), individuals are either fully susceptible to the current strain, or fully immune. Infection with a strain provides partial cross-immunity to the next strain, in the sense that each infected individual has a chance $(1-\nu)$ of developing full immunity, or else remains fully susceptible to the next strain. Assume further that each wave is brought about by a new strain of the virus, and that infection confers temporary protection against reinfection for the duration of the wave. This is consistent with epidemiological evidence from McMahon & Robb (2020), who find that SARS-CoV2 provides short term immunity against reinfection. These assumptions (or an equivalent version) have been used before in the network modeling literature (Bansal et al., 2010; Romanescu and Deardon, 2017).

Denote by t_l and T_l the start and end times of epidemic wave l, so that

$t_1 < T_1 < t_2 < T_2 < \ldots$. Notice first that the susceptible fraction at the start of wave $(l+1)$ is

$$S_{t_{l+1}} = S_{T_l} + \nu(1 - S_{T_l}) \tag{6}$$

i.e., the fraction is made up of individuals who remained susceptible at the end of the previous wave (S_{T_l}) plus a fraction of the previously infected who did not develop immunity to the newly arrived strain. Next, network models recognize that previously infected individuals tend to be more well connected than individuals who were never infected, as the chance of infection is proportional to the number of connections. By writing a balance equation of susceptibles for each degree k, we can demonstrate that the degree distribution of susceptible nodes at the start of a new wave (t_{l+1}) is a mixture of the degree distribution of the susceptible set at the end of the previous wave (group A), and the original distribution (group B), with weights $w_A = \frac{S_{T_l}(1-\nu)}{S_{t_{l+1}}}$ and $w_B = \frac{\nu}{S_{t_{l+1}}}$, respectively (see Appendix). In terms of the transmission dynamics, this is conceptually similar to splitting the susceptible population into two sub-networks having different degree distributions that spread infection at different rates.

An exact solution exists for R_t in terms of the original degree distribution, however this requires numerical methods, in addition to an assumption of said distribution (see Appendix). A simpler approach is to approximate the contact rate of the susceptible population by the weighted effective contact rates of the two sub-networks, weighted by the relative size of these sub-networks (w_A and w_B):

$$c_{l+1}\left(\frac{S_t}{S_{t_{l+1}}}\right) = w_A c\left(S_{T_l} \times \frac{S_t}{S_{t_{l+1}}}\right) + w_B c\left(\frac{S_t}{S_{t_{l+1}}}\right). \tag{7}$$

Notice that the susceptible fraction of the non-immune population in wave $l+1$, namely $\frac{S_t}{S_{t_{l+1}}}$, starts from 1 at t_{l+1}, and is found by dividing the susceptible fraction in the entire population (S_t) by those non-immune at the start of the wave ($S_{t_{l+1}}$). We then proceed to calculate the reproductive number as $R_{t+1} = \alpha S_t c_{l+1}\left(\frac{S_t}{S_{t_{l+1}}}\right)$, and $y_{t+1} = \alpha S_t c_{l+1}\left(\frac{S_t}{S_{t_{l+1}}}\right)\Lambda_t$. Finally, we should mention that a similar derivation may be arrived at without using networks, e.g., by assuming well-mixed ODE models with heterogeneous susceptibility among individuals. However, we have not found any similar treatment in the ODE literature that considers how the distribution of heterogeneous susceptibility changes in S_t following a loss of immunity among removed individuals.

Updates to S_t and Λ_t are made as before, and parameter ν will be estimated as part of the likelihood function. During a wave, one can use either actual observed values of y_t or predicted values when updating S_t, Λ_t and y_{t+1}. For retrospective model fitting we use predicted values of y_t, as this approach is consistent with the way ODE models are typically fitted, i.e., for a set of parameters the full curves are produced (agnostic of the observations), then a loss function is computed between the actual epidemic curve and the model curve. This approach may also converge faster, as for parameter settings far from the MLE, errors accumulate faster when using predicted values, producing visibly divergent curves. For predictions we should use all available information, and so we update using actual values. One important advantage to mention over ODE models, is that this model is self-priming, meaning that it does not require initial conditions to be estimated, hence reducing the number of necessary parameters to be fitted.

3. Results

3.1. Implementation details and settings

We fit the model to COVID-19 count data from two locations: the state of New York, and the city of Winnipeg. New York has been studied in our previous work on influenza (Romanescu and Deardon, 2017), and was found to be best described by relatively heavy-tailed distributions for the number of contacts of individuals. Although SARS-CoV2 and influenza are distinct respiratory diseases, with different transmissibility and serial interval profiles, based on our factorization (3) and update Eq. (1) the contact rate $c(\bullet)$ should be independent of these epidemiological factors. Winnipeg is a much smaller, provincial city in the Canadian Prairies, which stretches horizontally over a wide area, and where we would expect more homogeneity in transmission.

For the New York case data (from the New York Times Covid-19 dataset, 2022), we define the epidemic waves based on the start and end of the local decay pattern of empirical R_t, as outlined in subsection 2.1. We identify four waves up to our cutoff of March 1st 2022, as shown in Fig. S1, in the Supplementary Material. The date ranges are shown in Table 2. By searching through public health orders, we identify restrictions with varying levels of severity (coded from 0 to 3) that have been introduced at the times listed in Table 1. Note that severity levels are not standardized and it is up to the user to decide how many levels to use and how to define the thresholds. Qualitatively, we judged higher levels of restrictions to be associated with closure of schools, workplaces and entertainment venues, and lower levels (towards reopening) associated with capacity limits being imposed for venues. For the Winnipeg data (Manitoba Health website), a similar table is given in the Supplementary Material (Table S1). For the analysis of both datasets, we assume the serial interval (w_s) for COVID-19 from Nishiura et al. (2020) who model it as log-normal with median 4.0 days and standard deviation 2.9 days; we cap this at $M = 20$ days. The under-reporting rate (ρ_t) is informed from Irons and Raftery (2021) who compute the undercount fraction (which is $1/\rho$) of 2.1; so we assume a constant value for ρ of 1/2.1 or 47.6% for NY. There is no similar fraction computed for Winnipeg, so we use the value for Oregon from the same source, which gives $\rho = 1/2.2$, or 45.5%. This state was chosen because it is most similar with Manitoba in terms of largest city sizes (Portland vs. Winnipeg), average population density as well as cumulative rate of COVID-19 infections up to March 2022. The population sizes for the state of NY and the Winnipeg regional health authority are taken from the US Census Bureau and from the Government of Manitoba, respectively.

3.2. Fit to the full dataset and implied R_t curve

We first fit our model to the full time series of infection counts, for all shapes in Table 1. Parameter estimates for each contact rate curve $c(\bullet)$ and the mass action model are given in Table 3. Note that for the mass action model, ψ should be interpreted as a constant multiplier to the transmissibility for the duration of the restriction. For the variance parameters u and v in formula (4), it was found that including both, or only one did not make a noticeable difference in the maximum likelihood, so we set $v = -1$ and only estimate u.

The fitted curves are given in Fig. 1. As can be seen from the plot, the SNMs (which have a flexible R_t formulation) tend to fit observed data better than the mass action model, and this is confirmed by the AIC in

Table 2
Wave definitions and severity levels of restriction on the social network for the state of NY.

COVID-19 wave	Dates	Restriction severity level	Notes
Wave 1	March 18 – May 12, 2020	0, 3	3-lockdown introduced Mar 22
Wave 2	November 11 – June 2, 2021	1,2	1-limited reopening Sep 30 2-tightened restrictions Nov 21 1- limited reopening Feb 11
Wave 3	July 17 – October 10, 2021	0	0- restrictions lifted Jun 15
Wave 4	December 19, 2021 – January 30, 2022	0	

Table 3. The top performing shapes, based on the AIC, are the Granich, shifted inverse, followed by the two piece exponential. In particular, these shapes reach their peak in wave 3, whereas the other curves can only accommodate the increasing part of this wave. It is also encouraging to see that for these three shapes estimates of other parameters, such as relative transmissibility of waves, and restriction factors, are relatively consistent. For Winnipeg, the fitted curves are given in Fig. S2 in the Supplementary Material. Again, the top performers are the shifted inverse, Granich, and two piece exponential (in this order), and practically all SNMs outperform the mass action model (per the AIC values in Table S2).

We plot all the implied R_t curves in Fig. 2, for both locations. The differences between the two cities are pronounced; values of R_t are much higher overall in New York. There is some consistency in overall location and shape of the top three curves, which means that all SNMs agree about a curved/kinked R_t for both these populations. Note that for Winnipeg, the R_t implied by the Granich, Power and Geometric models are lower across the board, but this is because transmissibility estimates for waves 2 and 3 are high for these models compared to the rest. What is perhaps surprising is how fast R_t drops at the beginning of an epidemic in Winnipeg for two of the top three fitting curves. This suggests that there is perhaps a small, "connected core" of the city (which could be made up of essential workers, for instance), and once the epidemic clears this core, it will quickly run out of steam. It may also suggest an unusually high degree of compliance with public health orders, and a reluctance on the part of large segments of the population to abandon caution and to return to pre-pandemic activity levels, even after being given the green light by public health authorities.

3.3. Prediction study

A prediction study is carried out for each wave in the New York dataset separately, using the three best fitting shapes. Namely, for each day in a wave, the model is fit using all available information at the time, and the epidemic curve for the rest of the wave is predicted. Prediction quality is assessed as the mean squared prediction error (MSPE) of the future curve, averaged over all days in a wave. Predictions for the different shapes of $c(\bullet)$ are compared against one another and with a baseline mass action model which assumes a constant $c(S_t)$. It is important to note that due to automation of this prediction study, fits could not be inspected visually at each time point. Thus, better fits and predictions may be possible by using different optimizer settings at each time point, and the errors presented here are likely greater than what the models would produce in actual practice. To alleviate this problem, we start optimization for each day from a few different initial settings, and report the best prediction error (smallest MSPE) for each wave. Results are summarized in Table 4. The average is over all prediction days and all days in the future curves starting from each prediction day.

Among the four waves, the first and last were generally least well predicted, in absolute terms. This might be due to fewer observations being available in the first wave, making the models over-parameterized, and much higher nominal case counts in the last one, making absolute errors larger. All three shapes considered performed significantly better than the mass action model, and relatively similarly amongst themselves, on average. This is encouraging, and supports the idea that better fitting models tend to turn into better prediction models.

3.4. Assessing the effect of public health interventions

The framework developed can be used to study the effectiveness of public health interventions aimed at reducing social connections. This is captured via parameter(s) ψ; in general, ψ can be made to be time-varying and capture the real time connectivity of the network. For this illustration, we will only compare the three different levels of ψ corresponding to the severity levels in Table 2. Namely, we test the log likelihood difference to see if there is a significant improvement in quality of fit between a model with no restrictions ($\psi = 1$), and models with one, two, and three different restriction levels. Table 5 shows these differences sequentially, when adding one extra parameter. These differences multiplied by 2 are asymptotically chi-square distributed with one degree of freedom. Therefore, significance at the 5% level is given by an improvement in log likelihood of at least $3.84/2 = 1.92$.

Notice that all three shapes strongly support two differentiated levels of restrictions, but do not support adding a third level (ψ_3). Specifically, the data show a strong difference between unrestricted, limited reopening, and tightened restrictions, but does not find any added effect for lockdown. This could mean that periods of tight restrictions had a similar effect as full lockdown. Before leaning into these conclusions too

Table 3
Parameter estimates for NY state, using the full dataset. Maximum likelihood estimates and their standard deviations are shown. The (negative) log-likelihood value and the AIC are given at the bottom.

Parameters		Two piece exponential	Mass action	Shifted inverse	Poisson	Geometric	Granich	Power function
$\alpha \times c(S_t)$	a_0	–	3.29 (0.0167)	–	1 (–)	0.45 (0.065)	–	–
	a_1	1.30 (0.033)	–	0.58 (0.028)	2.72 (0.091)	4.09 (0.525)	5.05 (0.265)	3.5 (–)
	a_2	0.81 (0.004)	–	1.12 (–)	–	–	2.50 (0.061)	0.89 (0.059)
	a_3	5.56 (–)	–	–	–	–	0.59 (0.046)	–
	a_4		–	–	–	–	–	–
ν		1.00 (–)	0.00 (0.004)	0.56 (0.109)	0.00 (0.003)	0.00 (0.002)	0.863 (0.362)	0.00 (0.003)
relative variant transmissibility for waves 2 + ($v_1 = 1$)	v_2	1.26 (0.021)	0.348 (0.016)	1.22 (0.069)	0.99 (0.109)	0.45 (0.015)	1.26 (0.053)	0.36 (0.002)
	v_3	0.26 (–)	0.428 (–)	0.36 (0.043)	0.58 (0.021)	0.59 (0.019)	0.288 (0.071)	0.49 (0.006)
	v_4	0.40 (–)	0.550 (–)	0.56 (0.069)	0.82 (0.035)	1.06 (0.043)	0.449 (0.092)	0.76 (0.019)
restriction factors (ψ)	ψ_1	0.54 (0.019)	1.00 (–)	0.58 (0.019)	0.25 (0.039)	1.00 (–)	0.527 (0.021)	1.00 (–)
	ψ_2	0.49 (0.017)	1.00 (–)	0.52 (0.017)	0.23 (0.019)	0.90 (0.007)	0.476 (0.025)	1.00 (–)
	ψ_3	0.49 (0.017)	0.256 (–)	0.49 (0.042)	0.16 (0.037)	0.43 (0.022)	0.476 (0.021)	0.20 (0.023)
u		1129.41 (83.53)	3525.09 (287.807)	855.50 (63.36)	2392.02 (189.56)	1642.72 (125.16)	841.90 (63.66)	2522.57 (195.47)
$-l(\boldsymbol{\theta})$		3613.0	3840.46	3578.94	3761.91	3683.2	3576.99	3744.86
# of parameters		11	9	10	10	10	11	10
AIC		7248.0	7698.92	7177.88	7543.82	7386.4	7175.98	7509.72

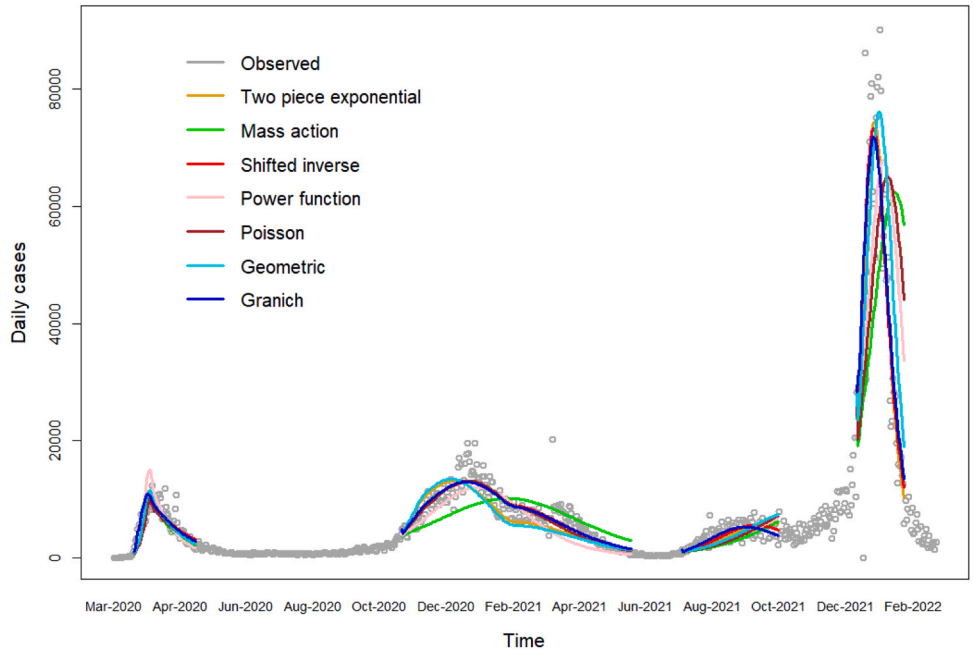

Fig. 1. Covid-19 infection data and model fits for NY. The fits are from the beginning to the end of the waves, as defined in the text.

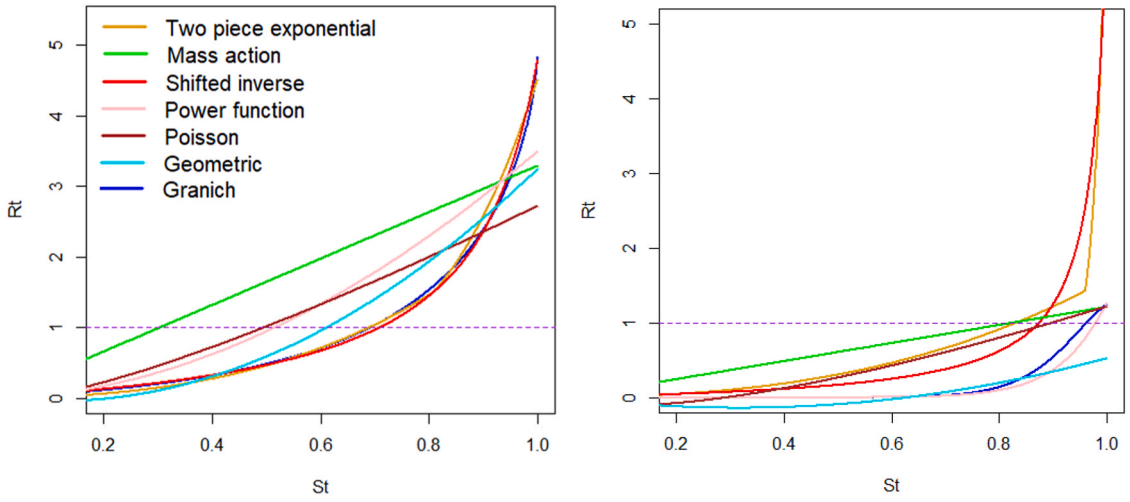

Fig. 2. R_t in an unrestricted network, over the range of S_t, as implied by the fitted models, for the NY dataset (left) and for Winnipeg (right). The endemic level ($R_t = 1$) is shown as a dashed line.

heavily, we should recognize the potential for confounding of several parameters in this analysis, especially when they act over the same time period. Lockdown measures are introduced in the first wave only, and, as such, their effect will be confounded by viral transmissibility particular to that wave. Although SNMs have some discriminatory power between the two effects, case data by itself affords limited insight into this matter. Thus, it is prudent to regard the analysis in Table 5 as hypothesis generating, and to seek confirmation from additional data sources, such as direct evidence of population mobility during this period.

Table 4
Results of prediction study for the New York data. The average represents the root of average squared deviations over all the days in the 4 waves.

Prediction Root MSPE (in '000)	Mass action	Granich	Two piece exponential	Shifted inverse
Wave 1	175.6	33.5	8.6	42.0
Wave 2	15.8	5.5	3.8	4.8
Wave 3	34.6	2.4	3.6	9.7
Wave 4	134.7	18.3	22.4	35.2
Average	**51.8**	**9.9**	**5.8**	**13.1**

Table 5
Differences in log likelihood when adding parameters phi sequentially, for each SNM. The unrestricted row gives the baseline level of negative log likelihood, to which differences are cumulatively added.

Restriction levels	Parametric conditions	Granich	Two piece exponential	Shifted inverse
unrestricted (baseline)	$1 = \psi_1 = \psi_2 = \psi_3$	3890.80	3868.09	3878.00
one level	$1 \geq \psi_1 = \psi_2 = \psi_3$	+ 281.94	+ 120.60	+ 237.70
two levels	$1 \geq \psi_1 \geq \psi_2 = \psi_3$	+ 31.39	+ 133.49	+ 61.30
three levels	$1 \geq \psi_1 \geq \psi_2 \geq \psi_3$	+ 0.48	+ 1.00	+ 0.06

4. Discussion

This paper introduced the synthetic network model, an integrated framework for prediction of infectious disease counts, based on the modeled time-varying effective reproductive number. While an SIR model based explicitly on R_t has been formulated recently (Abbott et al., 2020), attempting to model future trends in R_t is novel. We have also demonstrated its prediction performance in a retrospective study. There are a few key arguments why this R_t-based framework is preferable to the homogeneous mixing assumption, which is still widely used in most state-of-the-art models of infectious disease. To recapitulate:

1. The effective contact rate, as given by curve $c(S_t)$, is sufficient to capture incidence dynamics over any structured network (up to first order). Being a relatively stable feature of the population, the effective contact rate can be used to predict R_t at any stage of the epidemic (as measured by S_t), across waves, and for any pathogen. Essentially, the curve $c(\bullet)$ acts as a "synthetic network". Time-varying models of R_t do not have this feature, and they can only estimate R_t either retrospectively (Wu et al., 2020), or starting from a fully susceptible population (Chowell et al., 2016), which does not accommodate subsequent waves.
2. Formulating R_t as $\alpha \times S_t \times c(S_t)$ is interpretable and allows for differential modeling of public health interventions by distinguishing between measures to diminish p2p transmission, and measures to reduce (eliminate) social connections. These translate respectively into a reduction in α, and a reduction of the argument of $c(\bullet)$, i.e., making the last factor $c(\psi S_t)$. We are not aware of any other attempts to model this type of heterogeneous reduction in contacts.
3. Finally, we can plug $c(S_t)$ into traditional compartmental models as an adjustment factor for heterogeneity. Following the discussion in Section 2.1, $R_t = -(dS_t/dt)/I_t \times D$, from which we can obtain $dS_t/dt = -\alpha I_t S_t c(S_t)/D$. Thus, a standard compartmental model formulated as $dI_t/dt = bI_tS_t - \gamma I_t$ (where the last term describes recoveries) can be refitted by including the extra factor $c(S_t)$ into the first term, i.e., $\frac{dI_t}{dt} = bI_tS_tc(S_t) - \gamma I_t$. Note that since both α and D are constants, they are captured by the estimable parameter b, and so they do not need to be estimated separately. While this is not fully equivalent with our model because it does not account for the serial interval – ODE models usually do not – we expect that this formulation will provide the extra flexibility needed to fit data well, while being minimally invasive to existing models. This retrofitting can be done regardless of how complex the existing models is, or how many extra compartments it has, as there is usually only one step that defines new infections.

The implication of points 1 and 2 is more profound than just offering another way to relax the mass action assumption, and deserves further discussion in a broader modeling context. As approximations to first order network model dynamics, SNMs have a mechanistic basis (with the cautions outlined at the end of Section 2.1). In particular, they necessarily require the expression for the effective reproductive number to be $\alpha S_t c(S_t)$, which makes some previously proposed models inconsistent with a first order network model. For instance, Hethcote & van den Driessche (1991) assume a model for which the implied contact rate is $c(I_t) \propto I_t^{p-1}/(1 + bI_t^q)$. For this, and many other examples where a relaxation of the mass action assumption comes in conflict with the formulation in Eq. (3), there is no first-order network that could explain these contact rates for an SIR epidemic. Of course, this does not mean that there can't be another mechanistic explanation of disease transmission for these models. However, it does raise two important modeling questions: 1) how far away from a mechanistic basis is one comfortable working? and 2) are network models good mechanisms of disease spread? In cases such as prediction, the answer to the first question might be that a mechanism is irrelevant. However, practitioners will often want a model that can explain the spread patterns in more detail, or that can be used to answer what-if questions. In those cases, mechanistic models with the right resolution will surely be favored. Also, judging by the widespread use and persistence of the mass action assumption – itself mechanistic, – one would have to conclude that most users appreciate the mechanistic aspect of models. In answering the second question, network models are certainly appealing for describing the spread of a pathogen through a social network, and an individual's degree is a natural choice for describing heterogeneity in people's ability to transmit disease. Networks also offer extra flexibility to the modeler via assortativity (preferential mixing according to degree) and clustering. While we do not investigate these properties in the present paper, they have been studied in the theoretical modeling literature. Clustering has been modeled via the frequency of triangles in the network, and it was found that to account for clustering, one needs to multiply R_0 for the unclustered network by $(1 - \alpha C)$, where C is the clustering coefficient, although the impact of this adjustment was reported to be minor (Molina & Stone, 2012). Assortativity is thought to be more consequential to epidemic trajectory (Miller, 2009). We do not know how this might impact R_t, however, if the assortative mixing of the infectious set can be effectively indexed by the susceptible fraction, it may be possible to adapt Eq. (3) to cover this higher order behavior. While more research is warranted, there is a strong case to be made for networks as underpinning infectious disease transmission, and that SNMs are well positioned to link theoretical results to practice.

Computationally, SNMs implemented in a likelihood framework are very fast; the most complex model considered takes around one minute to fit to the full dataset on a laptop. This is significantly faster than similar implementations based on either deterministic curves fit via least squares, which require bootstrap or similar methods to produce error bounds for the parameters, or Bayesian implementations using Markov Chain Monte Carlo (MCMC). During a public health emergency when such models are needed, speed is certainly a benefit.

The main limitation of this approach is that due to the large number of parameters to be estimated, the MLE search routine can get trapped in local maxima in the likelihood surface. This means that different "optimal" fits may be found when starting from different initial values. This problem is by no means unique to these models; virtually all ODE models which are fit by optimizing some loss function will encounter the same problem. One mitigation strategy is starting the optimizer from two or a few different initial values, which increases the chances of finding the global maximum. A more radical solution lies in choosing more parsimonious models, as this problem is usually compounded by increasing complexity: a higher-dimensional parameter space increases the opportunity to have parameter combinations that produce similar likelihood values in different (distant) parts of the search space; this may create either ridges, or 'bubbles' of local maxima. While this may not be a major concern for obtaining good predictions, which tend to depend more on having a good fit rather than a realistic parameter set, caution is warranted when inferring individual parameters, and any results should be interpreted through the lens of expert opinion. Another limitation is that this model was presented in the SIRS framework, whereas COVID-19 data is often analyzed as SEIRS, which includes an intermediate exposed (E) compartment. For prediction purposes, having an exposed state is probably not essential, and we can think of the p2p transmissibility parameter α as the probability of transition through both states, i.e., $\alpha = P(S \to E) \times P(E \to I)$. However, if the application requires more specialized dynamics of the exposed state to be modeled, it is entirely possible to modify the present model to accommodate this complexity, in a similar way to how ODE models currently account for this (e.g., Tang et al., 2020).

A future direction of research will be relating the effective contact rate $c(S_t)$ to information about the distribution of contacts in the population, obtained from either survey data (Feehan and Mahmud, 2021) or cell phone mobility logs (Chang et al., 2021). Either of these could be used to inform the degree distribution of nodes in the network and we

plan to explore this in a follow-up paper. The benefit would be twofold. First, one would know in real time how the network of connections shifts in response to public policy, instead of having to wait for data to accumulate and estimate the effect retrospectively. This would result in better predictions, as has been demonstrated by using mobility data (Chang et al., 2021). Secondly, information about the network will allow an analyst to evaluate the effectiveness of public health interventions on disease spread. For instance, it will be possible to estimate what effect an x % reduction in the average number of contacts will have on new cases. The ability to perform calculations such as fine-tuning x (in this example) would be very valuable to public health officials by helping them contain a current epidemic without unduly disrupting social and economic contacts in the population.

Funding

This research received financial support from Research Manitoba, as part of its COVID-19 Research Fund. RGR is based at the George & Fay Yee Centre for Healthcare Innovation. Support for CHI is provided by University of Manitoba, Canadian Institutes for Health Research, Province of Manitoba, and Shared Health Manitoba.

CRediT authorship contribution statement

Razvan G. Romanescu developed the conception and design of the study, and created the models. Md Ashiqul Haque and Razvan G. Romanescu performed the computations and analysis of data. Md Ashiqul Haque, Razvan G. Romanescu, and Songdi Hu implemented the computer code and supporting algorithms. Mahmoud Torabi, Olivier Tremblay-Savard, Razvan G. Romanescu, and Douglas Nanton supervised the findings of the study. Douglas Nanton and Razvan G. Romanescu acquired the financial support for the project leading to this publication. Razvan G. Romanescu drafted the manuscript with revising support from all other authors.

Declaration of Competing Interest

We have no conflict of interests to disclose.

Appendix A

Mathematical network models

Definition of network quantities

Following Romanescu & Deardon (2017), assume that the population can be described by a fixed network with individuals as nodes, and connections as edges, and define the degree of a node (number of emanating edges) via random variable K with distribution $P(K=k)=p_k, k\geq 0$. Then, a basic result is that the chance of an individual becoming infected is proportional to its degree; explicitly, the survival probability of a susceptible with degree k at some time t is θ_t^k, for all $k\geq 0$ (Volz, 2008; Noël et al., 2009; Miller, 2011). Furthermore, θ_t is linked to the susceptible fraction via $g(\theta_t)=S_t$, where g is the probability generating function, or pgf of K, namely $g(x)=\sum_{k=0}^{\infty}p_k x^k$. To describe transmission dynamics, it is of interest to obtain the distributions of K_t^S and K_t^I, the degrees of the susceptible and the infectious sets at time t. Their pgf's are $g_t^S(x)=g(x\theta_t)/g(\theta_t)$, and $g_t^I(x)=g'(x\theta_t)/g'(\theta_t)$, respectively.

Dependence of R_t on the network

The definitions above allow us to derive the mean and variance of the expansion factor X_t, which is the number of secondary infections for one infected individual. Since $X_t|K_t^I\sim Binomial(n=K_t^I-1, p=\alpha S_t)$, where α is the per-edge probability of transmission over the entire infectious period, we arrive at

$$\begin{cases} E(X_t)=\alpha S_t E[K_t^I-1]=\alpha S_t g^{-1}(S_t)\dfrac{g'\left(g^{-1}(S_t)\right)}{g'\left(g^{-1}(S_t)\right)} \\ \mathrm{Var}(X_t)=E(X_t)\left(\alpha S_t g^{-1}(S_t)\dfrac{g''\left(g^{-1}(S_t)\right)}{g'\left(g^{-1}(S_t)\right)}+1-E(X_t)\right), \end{cases} \tag{1}$$

where g^{-1} is the inverse function to g, and g', g'' and g''' are the 1st, 2nd, and 3rd derivatives of g. Importantly, $E(X_t)=R_t$, the effective reproductive number.

As we are not likely to observe p_k or g directly, the take-aways are firstly that $R_t=\alpha S_t c(S_t)$, where $c()$ is the average contact rate of the infectious set (less one). Secondly, that c is a function of the network that evolves in time via S_t, in other words $c(t,S_t)=c(S_t)$, provided that the structure of the network doesn't change. Thus, S_t emerges as the natural "clock" of the epidemic. This observation has been made before in the similar but somewhat different context where heterogeneity refers to hosts having a varying probability to transmit a disease, given a contact. Novozhilov (2008) has shown that in this context, heterogeneity can be analyzed via a traditional SIR model with $dS_t/dt=-I_t h(S_t)$, where $h()$depends on the heterogeneity distribution, which is similar to the network result. As a note on language, we refer to c as the average contact rate of the infectious set for shorthand, but, in fact, it should be made clear that $c(S_t)=E(K_t^I)-1$, which reflects the fact that transmission can proceed to any of the connections except for the original infector. This is also referred to as excess degree (Molina and Stone, 2012).

Common network distributions

Three degree distributions found in the literature include:

- The Poisson distribution, $p_k=e^{-\lambda}\frac{\lambda^k}{k!}, k\geq 0$. The pgf is $g(x)=e^{\lambda(x-1)}$, and $R_t=\alpha S_t(\lambda+\ln(S_t))$.
- The geometric distribution: $p_k=\left(1-e^{-\frac{1}{\lambda}}\right)e^{-\frac{k}{\lambda}}, k\geq 0$, had pgf $g(x)=\frac{1-e^{-\frac{1}{\lambda}}}{1-xe^{-\frac{1}{\lambda}}}$. R_t simplifies to $R_t=2\alpha S_t\frac{e^{-\frac{1}{\lambda}}+S_t-1}{1-e^{-\frac{1}{\lambda}}}\approx 2\alpha\lambda S_t^2$, for large λ.

- The power law with an upper cutoff has $p_k \propto k^{-\lambda}$, for $k = 1, 2, \ldots, k_{\max}$. Here, the pgf is $g(x) = \sum_{k=1}^{k_{max}} \frac{x^k}{k^\lambda} / \sum_{k=1}^{k_{max}} \frac{1}{k^\lambda}$. In this case there is no closed form solution for R_t, however, the power law has the heaviest tail of the three distributions, which leads R_t to peak the highest (see Fig. 1 in Romanescu and Deardon (2017) for an example of a simulated epidemic and the corresponding R_t).

As can be seen in these common degree distributions, the function $c()$ is increasing, which makes the decay of R_t faster than under the mass action assumption.

Derivation of the degree distribution corresponding to an arbitrary contact rate $c(\bullet)$

If the contact rate function $c(S_t)$ is known, we may want to verify that there exists a counting variable K such that a first order network with degree K implies the contact rate function $c(\bullet)$. To do this, equate the two expressions for R_t and solve for the pgf g (note that $c(\bullet)$ means $\alpha c(\bullet)$, as this is what would be estimated from data):

$$xc(x) = \alpha x g^{-1}(x) \frac{g'(g^{-1}(x))}{g'(g^{-1}(x))}, \tag{2}$$

for any $x \in [S_{\min}, 1]$, where $S_{\min} < 1$ is the minimum value possible for S_t (also known as S_∞ in the literature, corresponding to the maximum size of the epidemic). Dividing by αx and integrating both sides leads to:

$$\frac{1}{\alpha}\int_{x_0}^{x} c(t)dt = \int_{x_0}^{x} g^{-1}(t)\frac{g'(g^{-1}(t))}{g'(g^{-1}(t))}dt = \int_{x_0}^{x} g^{-1}(t)\frac{dg'(g^{-1}(t))}{dt}dt = g^{-1}(x)g'\left(g^{-1}(x)\right) - C_0 - \int_{x_0}^{x} g'\left(g^{-1}(t)\right)\frac{dg^{-1}(t)}{dt}dt = g^{-1}(x)g'\left(g^{-1}(x)\right) - \int_{x_0}^{x} 1dt - C_0$$

$$= g^{-1}(x)g'\left(g^{-1}(x)\right) - x - (C_0 - x_0),$$

where we have used integration by parts. Next, if we let $h = g^{-1}$, we have $g'\left(g^{-1}(x)\right) = 1/h'(x)$, and the equality above can be written as

$$\frac{h(x)}{h'(x)} = \frac{1}{\alpha}\int c(x) + x + C_1 \tag{3}$$

where we use the notation $\int c(x)$ to mean the antiderivative of c, evaluated at x, and C_0 and C_1 are constants of integration. By taking the inverse of both sides in (1) and integrating again, we get

$$\ln(h(x)) = \int_{x_0}^{x} \frac{1}{\frac{1}{\alpha}\int c(t) + t + C_1}dt + C_2, \text{ or} h(x) = g^{-1}(x) = \exp\left\{\int_{x_0}^{x} \frac{1}{\frac{1}{\alpha}\int c(t) + t + C_1}dt + C_2\right\}. \tag{4}$$

Using the fact that $g(1) = h(1) = 1$, plugging $x = 1$ into (4) implies $C_2 = -\int \frac{1}{\frac{1}{\alpha}\int c(x)+x+C_1}\Big|_{x=1}$.

By inverting function $h(x)$ we obtain $g(x)$. To verify that g can indeed be the pgf of a random variable, differentiate it to obtain $p_k = \frac{1}{k!}g^{(k)}(0)$, for all $k = 0, 1, 2, \ldots$. If all p_k are nonnegative and sum to 1, then g is the pgf of random variable K with $P(K = k) = p_k$. Note that C_1 will need to be chosen to normalize the p_k. After a family of distributions with parameter(s) λ has been identified, one can again solve for the contact rate $c_\lambda(\bullet)$ which will now depend on λ. The model can now be parameterized in terms of λ instead of (a_1, a_2, a_3), giving the explicit dependence on the underlying degree distribution.

Derivation of R_t *under restrictions*

First, it is straightforward to show that weighting the probabilities p_k by an exponential tail results in the new probabilities $p_k^* = \frac{p_k\phi^k}{g(\phi)}$.

We find $g^*(x)$ as follows:

$$g^*(x) = \sum_{k=0}^{\infty} p_k^* x^k = \frac{1}{g(\phi)}\sum_{k=0}^{\infty} p_k x^k \phi^k$$

$= \frac{1}{g(\phi)}\sum_{k=0}^{\infty} p_k (x\phi)^k = \frac{g(x\phi)}{g(\phi)}$, by the definition of the p.g.f.

To find $g^{*-1}(x)$, let $x = g^*(y)$ and solve for y:

$$x = \frac{g(y\phi)}{g(\phi)} \Rightarrow g(y\phi) = xg(\phi) \Rightarrow y = \frac{g^{-1}(xg(\phi))}{\phi}$$

Therefore, $g^{*-1}(x) = \frac{g^{-1}(xg(\phi))}{\phi}$. Finally,

$$R_t^* = \alpha S_t g^{*-1}(S_t)\frac{g^{*'}(g^{*-1}(S_t))}{g^{*'}(g^{*-1}(S_t))}$$

$$= \alpha S_t \frac{g^{-1}(S_t g(\phi))\phi}{\phi}\frac{g'\left(\frac{g^{-1}(S_t g(\phi))}{\phi}\phi\right)}{g'\left(\frac{g^{-1}(S_t g(\phi))}{\phi}\phi\right)} \quad \left[\text{since } g^{*'}(x) = \frac{\phi g'(x\phi)}{g(\phi)} \text{ and } g^{*'}(x) = \frac{\phi^2 g'(x\phi)}{g(\phi)}\right]$$

$$= \alpha S_t g^{-1}(S_t g(\phi)) \frac{g^{'}(g^{-1}(S_t g(\phi)))}{g^{'}(g^{-1}(S_t g(\phi)))}$$

$$\therefore R_t^* = \alpha S_t c(S_t g(\phi)).$$

Likelihood function

The likelihood function for Section 2.4 with parameter vector $\boldsymbol{\theta}$ follows from distribution (5) in the text and is given by

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi \Lambda_t \alpha S_t c(\psi S_t)[u + (v+1)\alpha S_t c(\psi S_t)]}} \exp\left\{ -\frac{(y_{t+1} - \Lambda_t \alpha S_t c(\psi S_t))^2}{2\Lambda_t \alpha S_t c(\psi S_t)[u + (v+1)\alpha S_t c(\psi S_t)]} \right\}, \tag{5}$$

where $(y_1, y_2, \ldots, y_n)$ are the daily observed new infections.

Justification of SIRS formulas

The following derivation is based on the one in Romanescu and Deardon (2017). If $p_{k,t}$ is the probability that a random susceptible at time t has degree k, and the corresponding p.g.f. is g_t, then at time T_l we have $p_{k,T_l} = p_{k,t_l}\theta_l^k / g_{t_l}(\theta_l)$. The fraction of susceptible nodes of degree k at time t_{l+1} is $S_{t_{l+1}} N p_{k,t_{l+1}}$. It can also be written as the fraction of susceptibles remaining at time T_l, namely $S_{T_l} p_{k,T_l}$, and the fractions of degree k individuals that were previously immune but have again become susceptible to the new strain in wave $l+1$, and this is $\nu\left(p_{k,0} - p_{k,T_l} S_{T_l}\right)$. Also, since $\frac{S_{T_l}}{S_{t_l}} = g_{t_l}(\theta_l)$, we can put $p_{k,T_l} = p_{k,t_l}\theta_l^k S_{t_l} / S_{T_l}$. Thus, the balance equation at time t_{l+1} for susceptible nodes of degree k is:

$$S_{t_{l+1}} p_{k,t_{l+1}} = S_{T_l} p_{k,T_l} + \nu\left(p_{k,0} - p_{k,T_l} S_{T_l}\right).$$

Rearranging, $S_{t_{l+1}} p_{k,t_{l+1}} = S_{T_l} p_{k,T_l}(1-\nu) + \nu p_{k,0}$,

$$\Rightarrow p_{k,t_{l+1}} = \frac{S_{T_l}(1-\nu)}{S_{t_{l+1}}} p_{k,T_l} + \frac{\nu}{S_{t_{l+1}}} p_{k,0}, \text{ for all} k \geq 0. \tag{6}$$

Eq. (6) implies that, at the beginning of wave $l+1$, the degree distribution of susceptible nodes is a mixture of distributions p_{k,T_l} and $p_{k,0}$ with weights $w_A = \frac{S_{T_l}(1-\nu)}{S_{t_{l+1}}}$ and $w_B = \frac{\nu}{S_{t_{l+1}}}$. It can be verified that $w_A + w_B = 1$, using relation (6) in the main text. The p.g.f. of the mixture, $g_{t_{l+1}}(\bullet)$ can be written as

$$g_{t_{l+1}}(x) = \frac{S_{T_l}(1-\nu)}{S_{t_{l+1}}} g_{T_l}(x) + \frac{\nu}{S_{t_{l+1}}} g(x), \text{ for } all\ x \in [0,1],$$

where $g_{T_l}(x) = g(x\theta_t)/g(\theta_t)$, and $\theta_t = g^{-1}(S_{T_l})$. R_{t+1} is computed from Eq. (3) in the text, where the corresponding $c(\bullet)$ is computed using the mixture $g_{t_{l+1}}(\bullet)$ from above, namely

$$c_{t_{l+1}}(x) = {g_{t_{l+1}}}^{-1}(x) \frac{{g_{t_{l+1}}}^{'}\left({g_{t_{l+1}}}^{-1}(x)\right)}{{g_{t_{l+1}}}^{'}\left({g_{t_{l+1}}}^{-1}(x)\right)},$$

where x is the susceptible fraction for the non-immune population in wave $l+1$ (starting from 1 at t_{l+1}).

This formula requires a numerical procedure to invert the function $g_{t_{l+1}}(\bullet)$. A simplifying assumption would be to approximate $c_{t_{l+1}}(\bullet)$ as a weighted average of contact rates corresponding to the distributions p_{k,T_l} and $p_{k,0}$ with weights w_A and w_B. If we denote this approximation by $c_{t_{l+1}}(x)$ by $c_{l+1}(x)$, we have

$$c_{l+1}(x) = w_A c\left(S_{T_l} \bullet x\right) + w_B c(x).$$

Appendix B. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.epidem.2023.100708.

References

Abbott, S., Hellewell, J., Thompson, R.N., Sherratt, K., Gibbs, H.P., Bosse, N.I., Munday, J.D., Meakin, S., Doughty, E.L., Chun, J.Y., Chan, Y.W., 2020 1. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open. Research 5 (112), 112.

Anderson, S.C., Edwards, A.M., Yerlanov, M., Mulberry, N., Stockdale, J.E., Iyaniwura, S. A., Falcao, R.C., Otterstatter, M.C., Irvine, M.A., Janjua, N.Z., Coombs, D., 2020 3. Quantifying the impact of COVID-19 control measures using a Bayesian model of physical distancing. PLoS Comput. Biol. 16 (12), e1008274.

Bansal, S., Pourbohloul, B., Hupert, N., Grenfell, B., Meyers, L.A., 2010 26. The shifting demographic landscape of pandemic influenza. PloS One 5 (2), e9360.

Beutels, P., Shkedy, Z., Aerts, M., Van, Damme, P., 2006. Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface. Epidemiol. Infect. 134 (6), 1158–1166 (Dec).

Bjørnstad, O.N., Shea, K., Krzywinski, M., Altman, N., 2020. The SEIRS model for infectious disease dynamics. Nat. Methods 17 (6), 557–559. Jun 1.

Bowsher, C.G., Swain, P.S., 2012. Identifying sources of variation and the flow of information in biochemical networks. Proc. Natl. Acad. Sci. 109 (20), E1320-8.

Chang, S., Pierson, E., Koh, P.W., Gerardin, J., Redbird, B., Grusky, D., Leskovec, J., 2021. Mobility network models of COVID-19 explain inequities and inform reopening. Nature 589 (7840), 82–87 (Jan).

Chowell, G., Viboud, C., Simonsen, L., Moghadas, S.M., 2016. Characterizing the reproduction number of epidemics with early subexponential growth dynamics. J. R. Soc. Interface 13 (123), 20160659. Oct 31.

Connell R., Dawson P., Skvortsov A., Comparison of an agent-based model of disease propagation with the generalised SIR epidemic model, Technical report, DSTO, 2009.

Deardon, R., Brooks, S.P., Grenfell, B.T., Keeling, M.J., Tildesley, M.J., Savill, N.J., Shaw, D.J., Woolhouse, M.E.J., 2010. Inference for individual-level models of infectious diseases in large populations. Stat. Sin. 20 (1), 239–261.

Feehan, D.M., Mahmud, A.S., 2021. Quantifying population contact patterns in the United States during the COVID-19 pandemic. Nat. Commun. 12 (1), 1–9. Feb 9.

Gostic, K.M., McGough, L., Baskerville, E.B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J.A., De Salazar, P.M., Hellewell, J., 2020. Practical considerations for measuring the effective reproductive number, R t. PLoS Comput. Biol. 16 (12), e1008409. Dec 10.

Granich, R.M., Gilks, C.F., Dye, C., De Cock, K.M., Williams, B.G., 2009. Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. The Lancet 373 (9657), 48–57.

Hens, N., Calatayud, L., Kurkela, S., Tamme, T., Wallinga, J., 2012. Robust reconstruction and analysis of outbreak data: influenza A (H1N1) v transmission in a school-based population. Am. J. Epidemiol. 176 (3), 196–203. Aug 1.

Hethcote, H.W., van den Driessche, P., 1991. Some epidemiological models with nonlinear incidence. J. Math. Biol. 29 (3), 271–287 (Jan).

Hethcote H.W., Yorke J.A. Gonorrhea transmission dynamics and control. Springer; 2014 Mar 11.

Hornbeck, T., Naylor, D., Segre, A.M., Thomas, G., Herman, T., Polgreen, P.M., 2012. Using sensor networks to study the effect of peripatetic healthcare workers on the spread of hospital-associated infections. J. Infect. Dis. 206 (10), 1549–1557. Nov 15.

Huang, W., Cooke, K.L., Castillo-Chavez, C., 1992. Stability and bifurcation for a multiple-group model for the dynamics of HIV/AIDS transmission. SIAM J. Appl. Math. 52 (3), 835–854 (Jun).

Irons, N.J., Raftery, A.E., 2021. Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys. Proc. Natl. Acad. Sci. 118 (31), e2103272118. Aug 3.

Kermack, W.O., McKendrick, A.G., 1927. A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. Ser. A, Contain. Pap. A Math. Phys. Character 115 (772), 700–721. Aug 1.

Lau, M.S., Dalziel, B.D., Funk, S., McClelland, A., Tiffany, A., Riley, S., Metcalf, C.J., Grenfell, B.T., 2017. Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. Proc. Natl. Acad. Sci. 114 (9), 2337–2342. Feb 28.

Liljeros, F., Edling, C.R., Amaral, L.A., Stanley, H.E., Åberg, Y., 2001. The web of human sexual contacts. Nature 411 (6840), 907–908. Jun 21.

Machens, A., Gesualdo, F., Rizzo, C., Tozzi, A.E., Barrat, A., Cattuto, C., 2013. An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices. BMC Infect. Dis. 13 (1), 1–5 (Dec).

Miller, J.C., 2009. Percolation and epidemics in random clustered networks. Phys. Rev. E 80 (2), 020901. Aug 4.

Miller, J.C., 2011. A note on a paper by Erik Volz: SIR dynamics in random networks. J. Math. Biol. 62 (3), 349–358 (Mar).

Miller, J.C., Slim, A.C., Volz, E.M., 2012. Edge-based compartmental modelling for infectious disease spread. J. R. Soc. Interface 9 (70), 890–906. May 7.

Molina, C., Stone, L., 2012. Modelling the spread of diseases in clustered networks. J. Theor. Biol. 315, 110–118. Dec 21.

Molloy, M., Reed, B., 1998. The size of the giant component of a random graph with a given degree sequence. Comb., Probab. Comput. 7 (3), 295–305 (Sep).

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G.S., Wallinga, J., Heijne, J., 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 5 (3), e74. Mar 25.

New York Times Covid-19 dataset. Accessed October 4, 2022. ⟨https://github.com/nytimes/covid-19-data/blob/master/us-states.csv⟩.

Newman, M.E., 2002. Spread of epidemic disease on networks. Phys. Rev. E 66 (1), 016128. Jul 26.

Nishiura, H., Chowell, G., 2009. The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In Mathematical and statistical estimation approaches in epidemiology. Springer,, Dordrecht, pp. 103–121.

Noël, P.A., Davoudi, B., Brunham, R.C., Dubé, L.J., Pourbohloul, B., 2009. Time evolution of epidemic disease on finite and infinite networks. Phys. Rev. E 79 (2), 026101. Feb 2.

Petersen, E., Koopmans, M., Go, U., Hamer, D.H., Petrosillo, N., Castelli, F., Storgaard, M., Al Khalili, S., Simonsen, L., 2020. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. Lancet Infect. Dis. 20 (9), e238–e244. Sep 1.

Read, J.M., Eames, K.T., Edmunds, W.J., 2008. Dynamic social networks and the implications for the spread of infectious disease. J. R. Soc. Interface 5 (26), 1001–1007. Sep 6.

Romanescu, R.G., Deardon, R., 2017. Fast inference for network models of infectious disease spread. Scand. J. Stat. 44 (3), 666-83.

Romanescu, R.G., Deardon, R., 2019. Implementation of power law network models of epidemic surveillance data for better evaluation of outbreak detection alarms. Stat. Commun. Infect. Dis. 12 (1), 20180004. Jun 22.

Smith D., Moore L. The SIR model for spread of disease-the differential equation model. Convergence. 2004 Dec.

Stack, J.C., Bansal, S., Kumar, V.A., Grenfell, B., 2013. Inferring population-level contact heterogeneity from common epidemic data. J. R. Soc. Interface 10 (78), 20120578. Jan 6.

Stroud, P.D., Sydoriak, S.J., Riese, J.M., Smith, J.P., Mniszewski, S.M., Romero, P.R., 2006. Semi-empirical power-law scaling of new infection rate to model epidemic dynamics with inhomogeneous mixing. Math. Biosci. 203 (2), 301–318. Oct 1.

Tang B., Scarabel F., Bragazzi N.L., McCarthy Z., Glazer M., Xiao Y., Heffernan J.M., Asgary A., Ogden N.H., Wu J. De-escalation by reversing the escalation with a stronger synergistic package of contact tracing, quarantine, isolation and personal protection: feasibility of preventing a COVID-19 rebound in Ontario, Canada, as a case study. Biology. 2020 May 16;9(5):100.

Thompson, R.N., Stockwin, J.E., van Gaalen, R.D., Polonsky, J.A., Kamvar, Z.N., Demarsh, P.A., Dahlqwist, E., Li, S., Miguel, E., Jombart, T., Lessler, J., 2019. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics 29, 100356. Dec 1.

Van den Driessche, P., Watmough, J., 2002. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Math. Biosci. 180 (1–2), 29–48. Nov 1.

Volz, E., 2008. SIR dynamics in random networks with heterogeneous connectivity. J. Math. Biol. 56 (3), 293–310 (Mar).

Wallinga, J., Teunis, P., Kretzschmar, M., 2006. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. Am. J. Epidemiol. 164 (10), 936–944. Nov 15.

Wang, L., Didelot, X., Yang, J., Wong, G., Shi, Y., Liu, W., Gao, G.F., Bi, Y., 2020. Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. Nat. Commun. 11 (1), 1–6. Oct 6.

Webster, R.G., Bean, W.J., Gorman, O.T., Chambers, T.M., Kawaoka, Y., 1992. Evolution and ecology of influenza A viruses. Microbiol. Rev. 56, 152–179.

Wu S.L., Mertens A.N., Crider Y.S., Nguyen A., Pokpongkiat N.N., Djajadi S., Seth A., Hsiang M.S., Colford J.M., Reingold A., Arnold B.F. Substantial underestimation of SARS-CoV-2 infection in the United States. Nature communications. 2020 Sep 9;11 (1):1–0.