INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Okolie A, Müller J, Kretzschmar M. 2023 Parameter estimation for contact tracing in graph-based models. *J. R. Soc. Interface* **20**: 20230409. https://doi.org/10.1098/rsif.2023.0409

Received: 18 July 2023
Accepted: 1 November 2023

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
biomathematics, computational biology

Keywords:
stochastic susceptible–infected–recovered model on graph, contact tracing, epidemiology, branching process, parameter inference

Author for correspondence:
Augustine Okolie
e-mail: augustine.okolie@tum.de

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6927385.

THE ROYAL SOCIETY PUBLISHING

Parameter estimation for contact tracing in graph-based models

Augustine Okolie[1], Johannes Müller[1,2] and Mirjam Kretzschmar[3]

[1]Center for Mathematical Sciences, Technische Universität München, 85748 Garching, Germany
[2]Institute for Computational Biology, Helmholtz Center Munich, 85764 Neuherberg, Germany
[3]University Medical Center Utrecht, Utrecht University, 3584CX Utrecht, The Netherlands

AO, 0000-0002-8684-1358; JM, 0000-0001-5892-8598; MK, 0000-0002-4394-7697

We adopt a maximum-likelihood framework to estimate parameters of a stochastic susceptible–infected–recovered (SIR) model with contact tracing on a rooted random tree. Given the number of detectees per index case, our estimator allows to determine the degree distribution of the random tree as well as the tracing probability. Since we do not discover all infectees via contact tracing, this estimation is non-trivial. To keep things simple and stable, we develop an approximation suited for realistic situations (contract tracing probability small, or the probability for the detection of index cases small). In this approximation, the only epidemiological parameter entering the estimator is R_0. The estimator is tested in a simulation study and is furthermore applied to COVID-19 contact tracing data from India. The simulation study underlines the efficiency of the method. For the empirical COVID-19 data, we compare different degree distributions and perform a sensitivity analysis. We find that particularly a power-law and a negative binomial degree distribution fit the data well and that the tracing probability is rather large. The sensitivity analysis shows no strong dependency of the estimates on the reproduction number. Finally, we discuss the relevance of our findings.

1. Introduction

Infectious disease models have been instrumental in the study of many infectious diseases. Usually, these models are dependent on several biological parameters which can be epidemiological such as transmission, recovery, etc., or intervention parameters such as contact tracing, screening, vaccination and others. However, most of these parameters are not or only partially known and may cause predictions from these models to lack robustness [1] if not chosen appropriately. Missing data poses a major quantification challenge in epidemiology due to unobserved or partially observed events [2]. This makes parameter estimation essential in modelling disease spread. Often, the likelihood of parameters is maximized following the model predictions on sets of parameter values. In order to achieve parameter estimation, the model system property must be identifiable, i.e. estimating its parameters uniquely from the given data [3–5]. Several estimation methods, e.g. statistically based techniques such as approximate Bayesian computation (ABC) [6], Markov chain Monte Carlo (MCMC) integration [7], optimal control theory approach [8], classical least-squares method [9] and others (also see the review article [10]) have been instrumental in estimating parameters and making inferences in epidemic models. With respect to parameter estimation, contact tracing is particularly challenging as we somehow need to estimate the fraction of contacts we miss to identify: We need to estimate something that is *per definitionem* unobserved.

Several estimation techniques have been proposed for estimating important intervention parameters in modelling the recent COVID-19 pandemic. For instance, Manou-Abi *et al*. [11] obtained a best-fit model by proposing statistical methods for the underlying serial interval probability distribution for the COVID-19 virus in Mayotte from March 2020 to January 2022. Their method was then used to estimate time-varying reproduction numbers and transmission rates observed from the collected data.

Only a few attempts have been made to identify parameters specific to contact tracing: Müller & Hösel [12] proposed a branching process approach for contact tracing in randomly mixing populations to estimate tracing probability from contact history at the onset of an epidemic, based on the theory introduced in Müller *et al.* [13]. The derived estimator was then applied to data from contact tracing for tuberculosis and chlamydia. Blum & Tran [6] use a Bayesian framework to estimate parameters for rates of contact tracing and detection by random screening, and the method of Dyson *et al.* [14] is based on fitting a yaws and trachoma contact tracing survey data to a stochastic household model. Tanaka *et al.* [15] took up the branching process approach to estimate the percentage of undiagnosed persons in the COVID-19 pandemic with recursive full tracing.

In this paper, we propose methods for estimating parameters in graph-based models [16]. A stochastic susceptible–infected–recovered (SIR) model on a tree-shaped contact graph is modelled such that the underlying contact structure is given by a fixed or random graph. Due to the nature of the problem, we adapt the branching process theory results for contact tracing on random trees [16] to formulate a likelihood estimator for estimating the tracing probability and expected number of contacts. We first performed a simulation study with a Poisson degree distribution to check the performance of the maximum-likelihood estimator. Thereafter, we applied the model to contact tracing data collected during the COVID-19 pandemic in Karnataka, India. Overall, we show that our estimator based on the branching process theory for contact tracing is well suited for estimating tracing probabilities and degree distribution of the underlying contact structure in tree-based models.

The remaining part of the paper is structured as follows: §2 outlines the tree model and model assumptions. Section 3 presents the distribution of ages since infection, while §4 discusses the distribution of detected cases from one index case. We set up a likelihood estimator for estimating the tracing probability and underlying contact structure using these results and simulated data in §5 followed by a sensitivity analysis in §6. Last, we discuss our findings in §7.

1.1. Related works on SARS-CoV-2 epidemic models

Over the past few years, there has been a growing body of research on the mathematical modelling of infectious diseases, particularly the SARS-CoV-2 virus. Bertacchini *et al.* [17] provided insights into the temporal spreading of the virus, examining key parameters that influence the rate of spread. The work of Chondros *et al.* [18] presented an integrated simulation framework for both the prevention and mitigation of pandemics caused by airborne pathogens, providing a comprehensive approach towards understanding the dynamics of airborne diseases.

Furthermore, Cuevas-Maraver *et al.* [19] studied lockdown measures and assessed their impact on the COVID-19 outbreak in Mexico using both single- and two-age-structured epidemic models. Kovacevic *et al.* [20], on the other hand, employed a distributed optimal control epidemiological model for understanding the COVID-19 pandemic, emphasizing the importance of coordinated control efforts in disease mitigation. Modi *et al.* [21] focused on the spread of COVID-19 in India using the susceptible–exposed–infected–recovered (SEIR) model, providing crucial insights into the potential dynamics of the virus in dense populations. Kevrekidis *et al.* [22] added a spatial dimension to the modelling of COVID-19, studying the outbreak in Greece and Andalusia using reaction–diffusion models.

These works, while providing valuable insights into the dynamics and control of the SARS-CoV-2 virus, largely emphasize temporal, spatial, or control aspects. Our contribution, in contrast, focuses on the intricacies of contact tracing in the context of a stochastic SIR model implemented on a rooted random tree. We emphasize the challenge posed by undetected cases and the non-trivial nature of parameter estimation in such models. By offering a unique perspective on parameter estimation and contact tracing, we aim to add depth to the existing literature and contribute to a more comprehensive understanding of disease spread dynamics as observed in the COVID-19 outbreak.

2. Model assumptions

For the convenience of the reader, we will first sketch the motivation and idea of the branching theory process for contact tracing on rooted random trees (tree model) in Okolie & Müller [16] for our estimation analysis. A contact network in most applications represents individuals as nodes and interaction links via edges. Interaction links are channels where individuals can have direct or indirect contact, e.g. family, school, work, etc. These contact networks are applicable and useful in analysing contact tracing because they hold information about individuals and their neighbours [23,24]. However, applying contact networks to infectious disease dynamics is not straightforward as it requires a detailed understanding of the underlying network structure, e.g. the degree distribution and correlation, clustering coefficients, and properties defined by the network topology.

Once we have a defined contact network with predefined nodes and interacting links, we have a contact graph. The basic idea is to describe an epidemic by constructing a simple contact graph that is a rooted tree where only the root node is infected at the onset of the epidemic. The choice of this tree contact graph is for mathematical convenience as trees are not appropriate to describe more complex interactions for natural contact graphs. However, from a microscopic level, we can gain a better understanding of the overall mechanism and functioning of larger and more complex graphs, as many graph models as the configuration model resemble locally a tree [25]. Then we assign independently on each edge connecting one infected and one uninfected node a probability of transmitting the disease. If we focus on edges that transmit the disease, then we have only the infection graph, which is a subgraph of the contact graph. Contact tracing is also analysed on this infection graph such that upon recovery of an index case, direct neighbours of this index case are also removed with some tracing probability. From the number of detected cases by an index case via contact tracing, it is possible to estimate the degree distribution of the underlying contact network and also the tracing probability.

On the rooted random tree (figure 1), the infection starts from the root node R and spreads downwards through the directed edges. Individual C which is a direct contact of the root node is infected and spreads the disease to the focal individual A. Furthermore, A also spreads the infection to B and

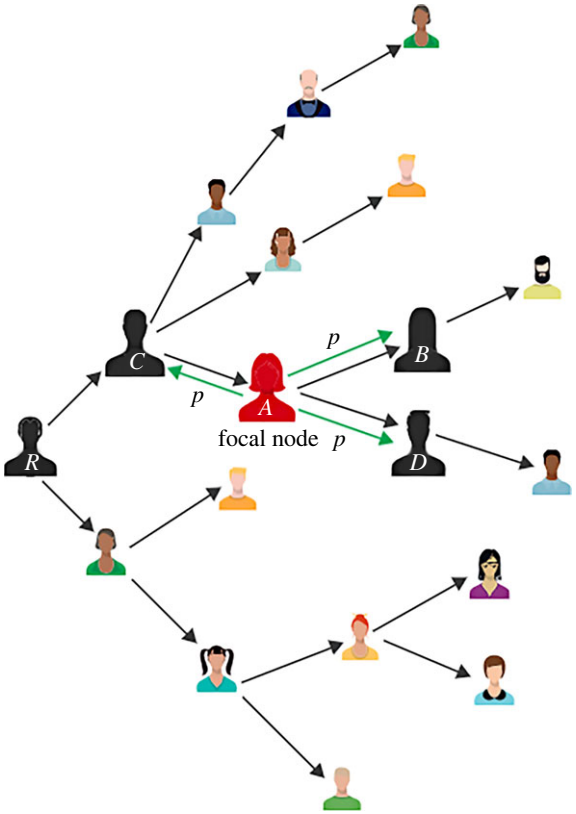

Figure 1. Schematic of the infection graph, illustrating the dynamics and interconnections in the process of disease transmission, and both forward and backward tracing. The root node is individual *R*. Focal individual *A*, infected by individual *C* (infector), subsequently infects individuals *B* and *D (infectees)*. In a forward tracing scenario starting from individual *A*, individuals *B* and *D* can be traced. By contrast, in a backward tracing scenario from *A*, individual *C* can be traced. The tracing probability is denoted in green.

D. Individuals *B* and *D* are the 'downstreams' (infectees) of *A* while *C* is an 'upstream' (infector) of *A*. We define *K* a random variable denoting the number of downstream edges of an individual with expectation $\mathbb{E}[K]$, where we assume that the downstream degree of each node is an independent and identically distributed (i.i.d.) realization of *K*. We note that the root node *R* is special as it has no infector, or equivalently, it has no upstream edge. It follows that the degree of the root node coincides with *K*, while all other nodes have a degree of $K+1$ (infectees plus infector). At the onset of the epidemic, only the root is infected while other individuals are susceptible. We consider a SIR model such that a recovered individual remains immune and does not get re-infected. Contact on one edge (between a susceptible and infected) will lead to infection. On a given edge, contacts happen at exponentially distributed waiting times at rate β. An infected individual recovers either unobserved at rate α, or observed and diagnosed at rate σ. Diagnosed individuals are immediately isolated or treated and classified as recovered. With probability $p_{\text{obs}} = \sigma/(\alpha+\sigma)$, an infected individual eventually is observed.

An observed/diagnosed infected individual not only becomes isolated but also is an index case that triggers contact tracing. That is, every adjacent edge has an independent probability p to be traced and consequently isolated if infected. In accordance with the data analysis we aim at, we focus here on one-step tracing, that is, traced individuals do not trigger further tracing events. We do, however, take into account forward and backward tracing as described in Okolie & Müller [16]. We do note this fact, as quite often, theoretical work solely focuses on forward tracing.

All in all, an infected individual can lose his or her infectivity in three possible ways; an unobserved recovery α, observed recovery σ and a successful tracing event. It turns out that the central ingredient for the analysis is the probability for an infected individual to still be infectious at age a. Please note that 'age' in the present paper always refers to the age of (or time since) infection, and never to chronological age. We define

$$\kappa(a) = \mathbb{P}(\text{a randomly chosen infected node of generation is infectious at age of infection } a), \tag{2.1}$$

which satisfies the following differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}a}\kappa(a) = -\kappa(a)(\alpha + \sigma + \text{tracing}(a)),$$

where $\kappa(0) = 1$. Without contact tracing, $\hat{\kappa}(a) := \mathrm{e}^{-(\alpha+\sigma)a}$. With contact tracing, this probability $\kappa(a)$ is decreased and thus

$$\kappa(a) = \hat{\kappa}(a)[1 - p \times \text{tracing in the interval } [0, a)]. \tag{2.2}$$

In [16], expressions for $\kappa(a)$ are derived. As we do not use these results in the current paper, we only indicate the overall structure and refer the interested reader to that paper for the details.

3. Distribution of ages since infection

In order to work out the distribution of the number of detectees per index case, the age since infection of the index case at its diagnosis is required. Thereto, we consider the case without contact tracing, $p = 0$ (such that index cases are diagnosed but do not trigger contact tracing). This assumption simplifies the arguments and yields an approximation for the age distribution in the case of $p > 0$, which is still appropriate if $p \ll 1$ or if $p_{\text{obs}} \ll 1$. It turns out, that the resulting approximation is sufficient for practical purposes.

Since the recovery rate α and the screening rate σ are constant, we have a Markovian model, and the age distribution of index cases coincides with the age distribution in the population.

Let $i(t, a)$ denote the age since infection-structured population size of infected individuals. As derived in Okolie & Müller [16], the age-structured model reads

$$(\partial_t + \partial_a) i(t, a) = -(\alpha + \sigma) i(t, a) \tag{3.1}$$

and

$$i(t, 0) = \int_0^\infty \theta(a)\, i(t, a)\, \mathrm{d}a, \tag{3.2}$$

where

$$\theta(a) = \mathbb{E}[K]\, \beta\, \mathrm{e}^{-\beta a}$$

is the age-dependent rate at which an infected average individual produces (downstream) infected. At this point, it is crucial that the contact graph is a tree, and the downstream degree distribution of each node/individual is an i.i.d. realization of the degree distribution *K*. If we count the number of nodes with a certain distance to the root, this number of nodes is exponentially increasing (for $\mathbb{E}[K] > 1$), unless the tree is finite in a given realization. We can exclude the case

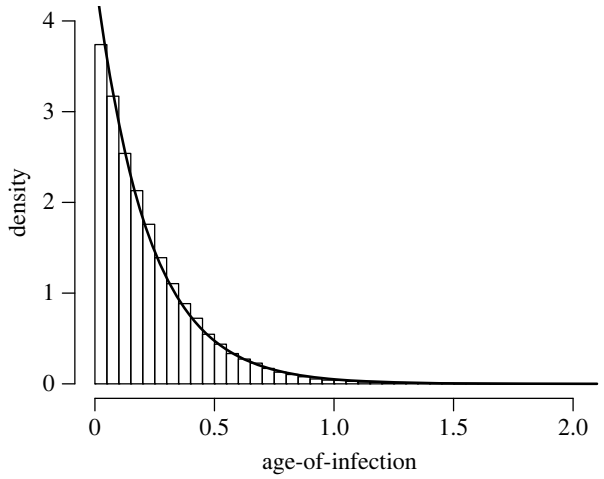

Figure 2. Theoretical age distribution $\varphi(a)$ (solid line) vs. simulated age distribution (bars). Parameters: $\beta = 1.5$, $\alpha = 0.5$, $\sigma = 0.5$, $p = 0.6$ and $\mathbb{E}[K] = 4$.

of finite trees since these realizations imply that we have a minor outbreak, and we are not interested in these minor outbreaks. As usual, the age-structured model will tend to an exponential growing solution with a stable age structure,

$$i(t, a) = I_0 \, \mathrm{e}^{\lambda t} \, i(a),$$

with $i(a) = \mathrm{e}^{-(\lambda+\alpha+\sigma)a}$ the probability to be infectious at age a. The exponent λ is the unique real root of

$$1 = \int_0^\infty \theta(a) \, \mathrm{e}^{-(\lambda+\alpha+\sigma)a} \, \mathrm{d}a = \mathbb{E}[K] \, \beta \int_0^\infty \mathrm{e}^{-(\lambda+\alpha+\sigma+\beta)a} \, \mathrm{d}a \Rightarrow \lambda$$
$$= \beta(\mathbb{E}[K] - 1) - \alpha - \sigma.$$

The asymptotic age distribution of index cases (which are detected at rate σ) tends to

$$\varphi(a) = \lim_{t\to\infty} \frac{\sigma \, i(t, a)}{\int_0^\infty \sigma \, i(t, b) \, \mathrm{d}b} = \beta(\mathbb{E}[K] - 1) \, \mathrm{e}^{-\beta(\mathbb{E}[K]-1)a}. \quad (3.3)$$

As a side remark, we also obtain the reproduction number from these considerations by

$$R_0 = \int_0^\infty \theta(a) \, \mathrm{e}^{-(\alpha+\sigma)a} \, \mathrm{d}a = \frac{\mathbb{E}[K] \, \beta}{\alpha + \sigma + \beta}.$$

As shown in figure 2, the agreement of the age distribution with simulated data is still excellent, though we have in the simulation $p = 0.6$ and $p_{\mathrm{obs}} = 1/2$. Furthermore, we have a higher density of lower age groups in the population. For any randomly chosen individual given by age since infection, it is not surprising to have a younger dominating age class. Due to the exponentially fast-growing population, this asymptotic age distribution is expected.

4. Distribution of detected cases

In this section, we derive the distribution of the number of detected cases per index case. That is the fraction of contacts of an index case who are detected via a tracing event triggered by the index case. We start with forward tracing. We then combine this result with backward tracing to yield full tracing.

Note that a central ingredient is the age distribution derived in the last section. We did not include contact tracing there. That is, all results in the present section are only a valid approximation if contact tracing does not crucially affect this age distribution. This is the case if either p or p_{obs} is small. All results are only valid under this assumption. However, the simulation study discussed below shows that this assumption is not too restrictive for practical purposes.

Proposition 4.1. *Let $\hat{p}(a)$ be the probability that an infected downstream node is successfully traced given that the focal individual becomes an index case at age since infection a.*

$$\hat{p}(a) = p \frac{\beta}{\alpha + \sigma - \beta} \left(\mathrm{e}^{-\beta a} - \mathrm{e}^{-(\alpha+\sigma)a} \right). \quad (4.1)$$

Proof. Note that an individual is only able to become an index case at the transition from I to R, that is, our focal individual is infectious in $[0, a)$. We consider one downstream individual. Let $s_1(c)$ represent the probability for this downstream individual to still be susceptible at age $c \in [0, a]$, $s_2(c)$ the probability to be infected, and $s_3(c)$ the probability for the downstream node to be removed (see figure 3). We have the following ordinary differential equations (ODEs):

$$\begin{aligned} \dot{s_1} &= -\beta s_1 & s_1(0) &= 1 \\ \dot{s_2} &= \beta s_1 - (\alpha + \sigma) s_2 & s_2(0) &= 0 \\ \dot{s_3} &= (\alpha + \sigma) s_2 & s_3(0) &= 0. \end{aligned}$$

The probability for the downstream node to be infectious at the time the infector has age since infection a given by $s_2(a)$,

$$s_2 = \frac{\beta}{\alpha + \sigma - \beta} \left(\mathrm{e}^{-\beta a} - \mathrm{e}^{-(\alpha+\sigma)a} \right),$$

and $\hat{p}(a) = p s_2(a)$ establishes the result. ∎

With this proposition and the age distribution $\varphi(a)$, we are able to find the distribution of the number of detected downstream individuals. For simplicity, we first consider a fixed degree distribution $K = k$ for some $k \in \mathbb{N}$, and then address the case of a random tree, where K is a random variable.

4.1. Fixed degree

In the present section, assume that the downstream degree of a node in the tree always is a deterministic number $K = k \in \mathbb{N}$. Particularly, $\mathbb{E}[K] = K$.

Proposition 4.2. *Let T be the random variable for the total number of successfully traced individuals by one index case and forward tracing only. The asymptotic probability distribution of T under forward tracing reads*

$$P(T = i) = \int_0^\infty \binom{k}{i} \hat{p}(a)^i (1 - \hat{p}(a))^{k-i} \beta(k-1) \, \mathrm{e}^{-\beta(k-1)a} \, \mathrm{d}a. \quad (4.2)$$

Proof. As we assume that a tracing event acts independently on different edges, the random variable T, conditioned on the age of the index case at diagnosis a, follows a binomial distribution with parameters k and an age-dependent tracing

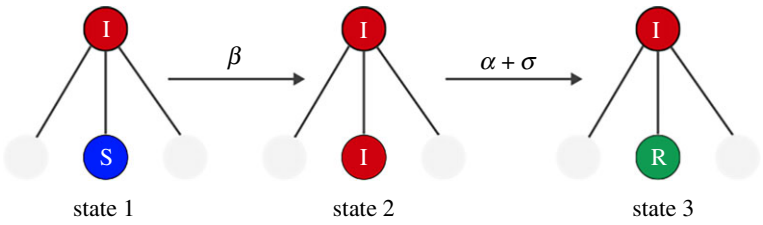

Figure 3. Transition states for a single edge in the infection process. These states represent the probabilities $s_1(c)$, $s_2(c)$ and $s_3(c)$ of an individual downstream from the index case to remain susceptible, infected and be removed at age $c \in [0, a]$.

probability on one edge $\hat{p}(a)$, $T \sim \text{Binom}(k, \hat{p}(a))$. Thus, the probability of i downstream detectees given age a and k total downstream nodes is given as

$$P(T = i \,|\, a) = \binom{k}{i} \hat{p}(a)^i (1 - \hat{p}(a))^{k-i}. \tag{4.3}$$

Last we remove the condition a by integrating over all possible age of index cases $\varphi(a)$ (equation (3.3)),

$$\begin{aligned} P(T = i) &= \int_0^\infty P(T = i \,|\, a)\, \varphi(a)\, \mathrm{d}a \\ &= \int_0^\infty \binom{k}{i} \hat{p}(a)^i\, (1 - \hat{p}(a))^{k-i} \beta(k-1)\, \mathrm{e}^{-\beta(k-1)a}\, \mathrm{d}a. \end{aligned} \tag{4.4}$$

■

Now we turn to full tracing. Thereto, we introduce the random characteristic I_a, which assumes the value 1 if the upstream individual of the index case (its infector) is still infected when the index case is identified (where the index case has age since infection a), and 0 else. Note that I_a is a Bernoulli random variable with

$$P(I_a = 1) = \mathrm{e}^{-(\alpha+\sigma)a} + \mathcal{O}(p).$$

As before, in what follows we use the approximation

$$P(I_a = 1) = \mathrm{e}^{-(\alpha+\sigma)a},$$

and drop the $\mathcal{O}(p)$ correction terms.

Proposition 4.3. *Let T_{tot} be the random variable for the total number of successfully traced individuals by one index case, under full tracing (forward and backward tracing). With $\varphi(a)$ and I_a as introduced above, the probability distribution of T_{tot} reads*

$$\begin{aligned} P(T_{\text{tot}} = i) = \int_0^\infty \Big[& pP(I_a = 1)\, P(T = i - 1 \,|\, a) \\ & + \big(1 - pP(I_a = 1)\big)\, P(T = i \,|\, a) \Big] \varphi(a)\, \mathrm{d}a. \end{aligned} \tag{4.5}$$

Proof. If the infector already is recovered ($I_a = 0$), then (conditioning on the age/time of infection of the index case a)

$$P(T_{\text{tot}} = i|a,\ I_a = 0) = P(T = i|a).$$

If the infector is still infectious, also the infector might be traced, such that one of the k detectees might be the upstream individual (probability p), or not (probability $1 - p$),

$$P(T_{\text{tot}} = i|a,\ I_a = 1) = p\, P(T = i - 1|a) + (1 - p)\, P(T = i|a).$$

Taking these two cases together, we have

$$\begin{aligned} P(T_{\text{tot}} = i|a) &= P(I_a = 0)\, P(T = i|a) + P(I_a = 1) \\ &\quad \times \Big(p\, P(T = i - 1|a) + (1 - p)\, P(T = i|a) \Big) \\ &= pP(I_a = 1)\, P(T = i - 1 \,|\, a) \\ &\quad + \big(1 - pP(I_a = 1)\big)\, P(T = i \,|\, a). \end{aligned}$$

Integrating by $\varphi(a)\, \mathrm{d}a$ removed the condition on a and yields the result. ■

4.2. Random degree

So far, the model is formulated for a fixed degree. In most applications, we do not always know individual contacts k due to randomness in contact structure. We now assume an arbitrary degree distribution such that the distribution of the contacts of a random node is defined by some probability distribution $P(K = k)$. The model for fixed case in equation (4.3) is adapted; we only have to take the expectation by summing over all possible numbers of contacts k multiplied by the corresponding probabilities. Thus,

$$\begin{aligned} P(T_{\text{tot}} = i) = \sum_{k=i}^\infty \int_0^\infty \Big[& pP(I_a = 1)\, P(T = i - 1 \,|\, a,\ K = k) \\ & + \big(1 - pP(I_a = 1)\big)\, P(T = i \,|\, a,\ K = k) \Big] \varphi(a)\, \mathrm{d}a\, P(K = k). \end{aligned} \tag{4.6}$$

As illustrated in figure 4, we have the distribution of the number of detected secondary cases via contact tracing. For the parameters chosen in our study (figure 5), we find a good agreement between our theory results and simulation. We again emphasize that the age structure entering our estimation is only an approximation, as contact tracing is neglected. Nevertheless, the results are more than acceptable, even for $p = 0.6$ and $p_{\text{obs}} = 0.5$.

We note that our estimator is independent of time t. The only 'time' that appears is the age since infection a. The probability $P(T_{\text{tot}} = i)$ consists of integrals as $\int_0^\infty g(a)\, \varphi(a)\, \mathrm{d}a$. Here, we are allowed to choose the time unit, resp. to define $a = \zeta b$ for $\zeta > 0$,

$$\int_0^\infty g(a)\, \varphi(a)\, \mathrm{d}a = \int_0^\infty g(b\,\zeta)\, \zeta \varphi(b\,\zeta)\, \mathrm{d}b.$$

If we choose one time unit to be $1/(\alpha + \sigma)$, which is $\zeta = 1/(\alpha + \sigma)$, then in $P(T_{\text{tot}} = i)$ the epidemiological parameters β, α and σ can always be replaced by an expression of R_0 and $E[K]$. That is, the epidemiological parameters enter the estimator

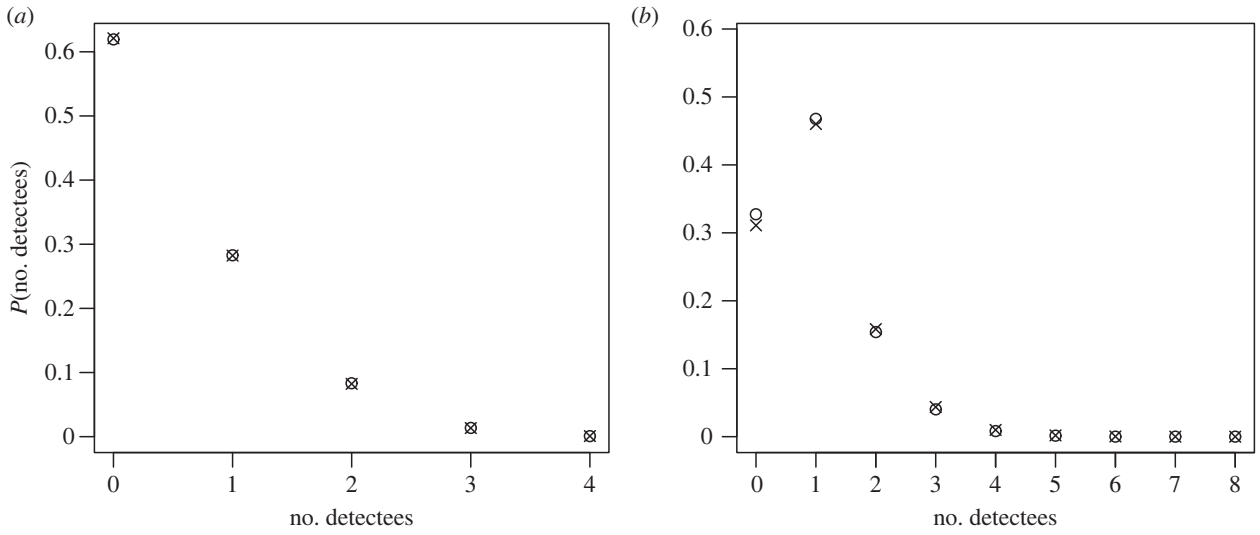

Figure 4. Distribution of detected cases per index case: theoretical predictions vs. simulated results. (*a*) Forward tracing for fixed degrees. (*b*) Full tracing for a Poisson graph. In both panels, crosses denote theoretical probabilities $P(T=i)$ or $P(T_{\text{tot}}=i)$, whereas circles represent results from 100 000 simulations. Additional parameters used during the simulation: $\beta=1.5$, $\alpha=0.5$, $\sigma=0.5$, $p=0.6$ and $\mathbb{E}[K]=4$.

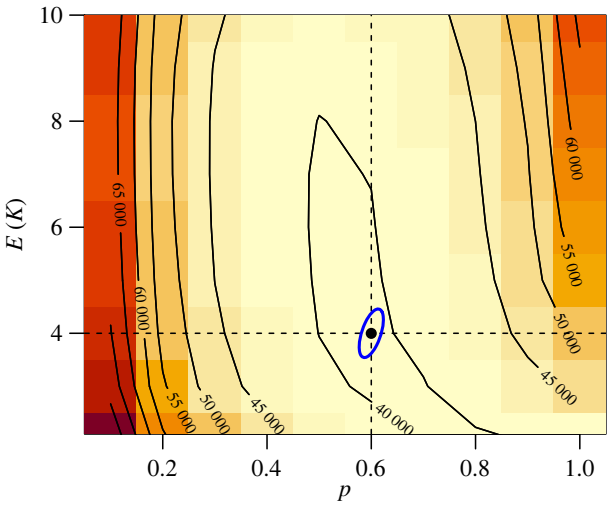

Figure 5. Parameters estimated from the simulation data derived from 100 000 iterations. The tracing probability, denoted as p, was determined to be 0.6 and the expected number of edges, represented by $\mathbb{E}[K]$, was 4. The blue region represents the 95% confidence region for the estimated parameters. Additional parameters used during the simulation: $\beta=1.5$, $\alpha=0.5$, $\sigma=0.5$.

solely via R_0. We only check that fact for one of the terms, as the argument is similar for the other terms.

$$\begin{aligned}
&\int_0^\infty P(I_a=1)\; P(T=i\,|\,a,K=k)\,\varphi(a)\,\mathrm{d}a\\
&=\binom{k}{i}\int_0^\infty \mathrm{e}^{-(\alpha+\sigma)a}\;\hat{p}(a)^i\,(1-\hat{p}(a))^{k-i}\;\beta(\mathbb{E}[K]-1)\,\mathrm{e}^{-\beta(\mathbb{E}[K]-1)a}\,\mathrm{d}a\\
&=\binom{k}{i}\int_0^\infty \mathrm{e}^{-b}\;\hat{p}\left(\frac{b}{(\alpha+\sigma)}\right)^i\left(1-\hat{p}\left(\frac{a}{(\alpha+\sigma)}\right)\right)^{k-i}\\
&\quad\times\frac{\beta(\mathbb{E}[K]-1)}{\alpha+\sigma}\,\mathrm{e}^{-(\beta(\mathbb{E}[K]-1)/(\alpha+\sigma))\,b}\,\mathrm{d}b.
\end{aligned}$$

With $R_0=\beta\,\mathbb{E}[K]/(\alpha+\sigma+\beta)$ we have (note that always $\mathbb{E}[K]>R_0$, as we have—on average—only $\mathbb{E}[K]$ downstream individuals who can get infected)

$$\frac{\beta}{\mu+\sigma}=\frac{R_0}{\mathbb{E}[K]-R_0},$$

and hence

$$\begin{aligned}
\frac{\beta\,(\mathbb{E}[K]-1)}{\alpha+\sigma}&=\frac{(\mathbb{E}[K]-1)\,R_0}{\mathbb{E}[K]-R_0},\\
\hat{p}\left(\frac{b}{(\alpha+\sigma)}\right)&=\frac{\beta/(\alpha+\sigma)}{(\beta/(\alpha+\sigma))-1}\,(\mathrm{e}^{-b}-\mathrm{e}^{-(\beta/(\alpha+\sigma))\,b})\\
&=\frac{R_0}{2\,R_0-\mathbb{E}[K]}\left(\mathrm{e}^{-b}-\mathrm{e}^{-(R_0/(\mathbb{E}[K]-R_0))\,b}\right).
\end{aligned}$$

That is, all expressions only depend on R_0, K and p.

Corollary 4.4. *$P(T_{tot}=i)$ and $P(T=i)$ only depend on the epidemiological parameters via R_0 and depends furthermore on the degree distribution given by K and on the tracing probability p.*

We can use the formulae from above, where we pragmatically set $\alpha+\sigma$ to 1, and—given R_0 and $\mathbb{E}[K]$—define $\beta=R_0/(\mathbb{E}[K]-R_0)$.

5. Estimation by maximum-likelihood method

In this section, we will set up the likelihood estimator for our model. We assume that we have n observations of index cases, and where $i_\ell\in\mathbb{N}_0$, $\ell=1,\ldots,n$ denote the total number of detectees per index case (one-step tracing only).

5.1. Likelihood estimator

We are able to set an estimator via $P(T_{\text{tot}}=i)=P(T_{\text{tot}}=i\mid\boldsymbol{\mu})$ for these data points where $\boldsymbol{\mu}$ are the parameters of the model we wish to estimate (tracing probability and parameters of the random variable K, e.g. $\mu=(p,\mathbb{E}[K])$ in the case of a Poisson distribution for K). The likelihood for the data reads

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\mu}\mid i_\ell,\ell=1,\ldots,n)=\prod_{\ell=1}^{n}\int_0^\infty\Big[&pP(I_a=1)\,P(T=i_l-1\,|\,a)\\
&+\big(1-pP(I_a=1)\big)\,P(T=i_l\,|\,a)\Big]\varphi(a)\,\mathrm{d}a,
\end{aligned}$$

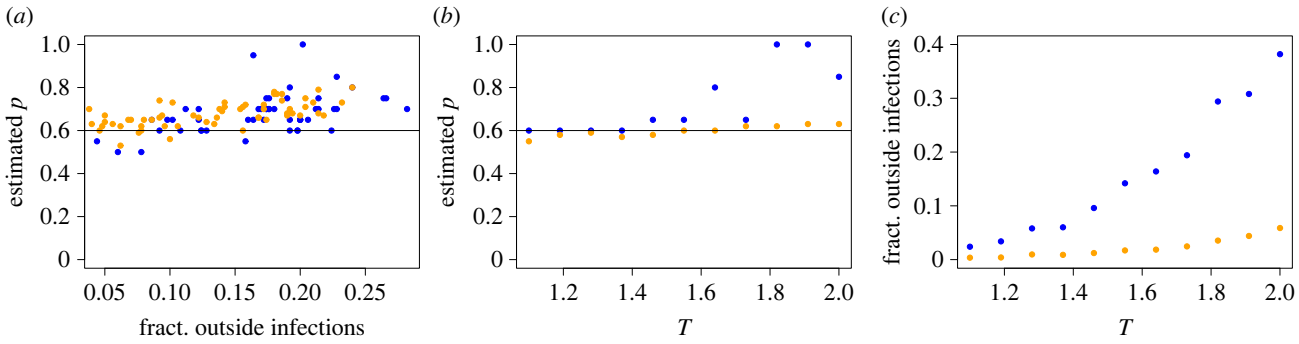

Figure 6. Results of the simulation study for the performance of the estimator for the generalized configuration model (orange: fixed excess degree, blue: Poisson excess degree). (*a*) Estimated p over the fraction of index cases with outside infections, induced by forcing the configuration graph to have triangles. For the fixed degree, index cases from the time interval [0, 2] are used, for the Poisson degree distribution, we use the time interval [0, 1.2]. (*b*,*c*) We consider the index cases identified in the time interval [0, T], where T is on the x-axis. The standard configuration model is used to produce the graph, without forcing for additional triangles. (*b*) Estimation of p. (*c*) Fraction of index cases with outside infections. Parameters: $\beta = 1.5$, $\alpha = \sigma = 0.5$, $p = 0.6$, $\mathbb{E}(K) = 4$, only forward tracing.

and the log-likelihood is given by

$$\mathcal{LL}(\boldsymbol{\mu} \mid i_\ell, \ell = 1, , \ldots, n) = \sum_{\ell=1}^{n} \ln\left(\int_0^\infty \Big[pP(I_a = 1)\, P(T = i_l - 1 \mid a) + \big(1 - pP(I_a = 1)\big)\, P(T = i_l \mid a) \Big] \varphi(a)\, \mathrm{d}a \right). \tag{5.1}$$

5.2. Simulated data

An agent-based stochastic model is used to simulate the data. To maximize the likelihood in equation (5.1), we plug into the likelihood function all independent observed data points and determine the arg max. As shown in figure 5, in our estimation we find back the true values we used in the simulation, namely the tracing probability $p = 0.6$ and the expected number of edges $\mathbb{E}[K] = 4$. The blue circle region contains a global maximum for the true parameter value of the estimated Poisson degree distribution. Other parameters β, α, σ are known and fixed. For the simulated data, we find a satisfying result based on our theory assumption.

5.2.1. Stability of the estimator against non-tree typologies in the contact graph

Real-world networks are, of course, no trees. We investigate the stability of our estimators against the violation of this prerequisite. Our estimator is particularly based on two assumptions: we know the distribution of the index cases' time since infection, and all downstream nodes are susceptible when a node becomes infected.

The notion of downstream nodes as introduced above depends on the tree topology. We generalize this notion in calling all neighbours of a focal infected node 'downstream nodes' apart from the infector of this focal node. Circles and clusters in a contact graph might lead to infections of downstream nodes from outside (which we call 'outside infections').

The deviation from a tree can be measured on the topological level or on an immunological level. Speaking about topology, particularly the appearance of triangles is well known to affect theory which is based on a tree topology: the message-passing method [26,27] is an exact version of the pair approximation on trees. Pair approximation, in turn, requires correction terms if considered on more general networks [24]. We thus expect that triangles might challenge our estimator. We use the configuration model [25] in an adapted version which allows to control the fraction of nodes in triangles (see electronic supplementary material).

The second level where the deviation from trees becomes visible is the epidemiological process: the fraction of outside infections is an alternative characterization for non-tree graphs. It turns out (see electronic supplementary material) that this epidemiological characterization better predicts the performance of the estimator than the density of triangles: outside infections slow down the spread of an epidemic, and in that, the time-since-infection structure of index cases is shifted to longer infectious periods. In that, index cases have more time to infect their downstream nodes. Moreover, outside infections produce potentially even more infected downstream nodes as we expected from tree-based models. Both effects point in the same direction, such that p tends to be overestimated to explain the additional detectees which are induced by the epidemiological consequences of the graph topology.

If we inspect the simulation study (figure 6*a*), we find that indeed the estimates of p increase over the fraction of index cases which possess outside infections, which are introduced by triangles. However, up to a fraction of 10–20%, this effect is not too severe.

The next, interesting question is the influence of the excess degree distribution instead of the influence of triangles on the fraction of outside infections. If we compare a configuration model with a fixed excess degree and a Poisson excess degree over a first time interval [0, T], we find in the Poisson graph a much faster increase of these outside infections in T. The difference between fixed and Poisson excess degree is much more important than the number of triangles (see electronic supplementary material). Indeed, for the Poisson excess degree, the estimator becomes biased even during the late exponential growing phase, while in the fixed degree model, the estimations are acceptable basically during the complete exponential growing phase (figure 6*b*); the reason is the striking difference in the number of index cases with outside infections (figure 6*c*). The explanation is well known: configuration models locally look like trees. However, as it is known from the celebrated friendship paradox [28–30], the preferential mixing of the configuration model lets nodes connect to nodes with a high degree. Therefore, we find in the case of the Poisson excess degree a

Table 1. Total number and frequencies of detected cases.

no. detectees	0	1	2	3	4	5	6	7	8	10	11	12	13	15	16	19	22	28	29
frequency	766	87	34	19	16	12	3	4	3	2	1	1	1	1	1	2	1	1	1

relatively clustered, small subgroup of nodes, where the infection will take place first. We have a kind of well-connected core group, which is distinctly smaller than the population size. Thus, outside infections are likely to take place after the infection moves into this core group. In the fixed degree, we of course cannot have such a core group, which explains the stability of the estimations in the fixed degree model.

In the long run, however, the assumptions of an exponentially growing prevalence will not be given any more, independent of the excess degree distribution. The prerequisites of our estimator are not valid any more. In section S4 of electronic supplementary material, we indicate a possibility of extending the basic ideas developed for trees, in order to also cover the long-term behaviour of an epidemic. However, this question is not the focus of this present work.

A central question now is the applicability of the theory to real-world data on the background of these simulation results. This might depend on the transmission mechanisms. Sexually transmitted diseases (STDs) are known to depend on core groups, and the contact network is less dynamic as in respiratory diseases. That is, we should be careful in applying our estimator to STDs. The infectious contacts of respiratory infections are known to exhibit over-dispersion [31]. However, as many of the contacts are rather casual, the infection network is not static and we will not find a fixed core group. In that, we expect that the tree-based estimators will work fine for respiratory infections.

5.3. COVID-19 data

In the previous section, we used the maximum-likelihood estimator with contact tracing on simulated data to estimate the tracing probability and expected number of edges. In the simulated dataset we analysed, we have information about the total number of contacts of index cases and also the number of infected contacts identified via contact tracing. In this section, we would like to see how this method works with empirical data. We wish to estimate the tracing probability p and expected number of edges $\mathbb{E}[K]$ from a dataset collected during the COVID-19 pandemic. We obtained a published dataset on contact tracing conducted in 2020 from a remarkable extensive and nice study in Karnataka, India where 956 cases with confirmed forward contact tracing were reported between the 9 March and the 20 May 2020 [32]. A comprehensive description of the dataset including the data source, data handling and ethics approval can be found in Gupta *et al.* [32]. A summary of the number of detectees per index case from the reported dataset is shown in table 1.

We only look at one-step tracing at the moment because only detected secondary cases of primary (index) cases are accounted for in the likelihood estimator. Generally, in most epidemic modelling studies, secondary cases are defined as close/direct contacts (e.g. household, family, etc.) of index cases [33]. In the dataset, direct and indirect detected contacts of index cases were reported. For our study, we would focus on only the direct contacts, thus we are able to analyse this scenario with the estimator for only one-step forward tracing. For the reported reproduction number, we chose $R_0 = 3$ in accordance with Gupta *et al.* [32]. An extensive meta-analysis of COVID-19 data from China encompassing 29 studies revealed an approximate R_0 value of 3.32 (95% CI: 2.81–3.82) [34]. In order to investigate the influence of R_0 on our estimates, we additionally carried out a sensitivity analysis.

Table 2. Examples of standard random graph models.

network model	degree distribution for large population size N
full graph/random mixing[a]	$K = N - 1 \to \infty$, $\beta \to 0$, R_0 constant
Erdös–Rényi	$K \sim$ Poisson
configuration model	choice: $K \sim$ geometric
scale-free network	power-law, $P(K = k) = ck^{-\gamma}$, $\gamma > 1$
standard degree distribution	negative binomial

[a]See appendices A and B for further detail on full graph/random mixing and the optimization process, respectively.

We use standard random graph models (table 2) to get inspiration on which degree distribution might be appropriate for describing the data [25].

We performed (i) a maximum-likelihood estimation and (ii) for model comparison, we used the Akaike information criterion (AIC) and a chi-square (goodness-of-fit) test. The summary of the point estimates is shown in table 3.

Optimization. We inspected the gradient of the result and the eigenvalues of the Hessian to ensure that we have (at least) an approximate local maximum. The estimator converged satisfyingly for all models except for the Poisson degree model: in that case, the parameter of the distribution (the expectation) always increased. Seemingly, the optimum is either very large or even infinite.

Confidence intervals. The approximate 95% confidence intervals are based on the quadratic approximation of the log-likelihood at its maximum, respectively, the approximation of the Fisher information matrix by the inverse of the negative Hessian. We determined the confidence intervals for geometric, scale-free and negative binomial distributions, as the other models turn out to be inappropriate for the data. In the case of the negative binomial distribution, however, the log-likelihood attains its maximum close to the theoretical lower boundary of $\mathbb{E}[K]$ which is R_0. As this is numerically a delicate situation, we approximated the confidence interval for $\mathbb{E}[K]$ in that case by the maximum value of $\mathbb{E}[K]$ such that

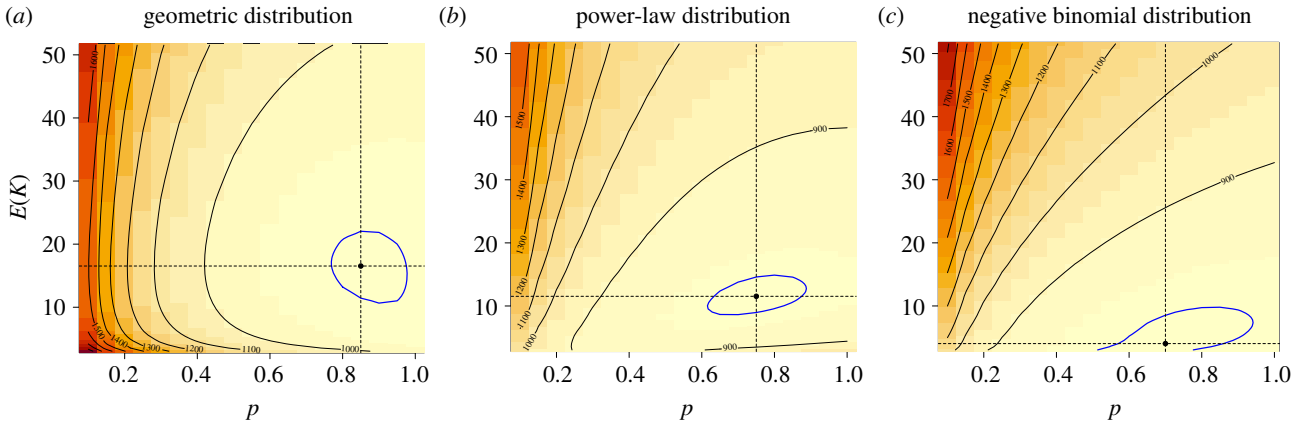

Figure 7. A contour plot showing the point estimates p and $E(K)$ of the likelihood estimator in different network models. Distributions from (*a*)–(*c*): Geometric, scale-free and negative binomial. Parameters; $R_0 = 3$, for the negative binomial, $r = 0.17$.

Table 3. Parameter estimates and model comparison for five probability distributions fitted to a dataset, including the probability (p), the expected value ($\mathbb{E}[K]$), additional information (if available), AIC and p-value from a chi-squared test. Note that the estimator in the case of the Poisson distribution did not converge, and we simply fixed a large expected value. The interval for r in the negative binomial distribution is 95% CI.

distribution	p (95% CI)	$\mathbb{E}[K]$ (95% CI)	add. information	AIC	p-value (χ^2)
random mixing	0.98	—		2443	$<10^{-20}$
Poisson	0.98	52	(not converged)	2464	$<10^{-20}$
geometric	0.87 (0.76, 0.97)	16.6 (11.3, 22.0)		1858	$<10^{-20}$
power-law	0.74 (0.61, 0.88)	11.4 (8.3, 14.5)	$\gamma = 1.48$	1687	0.003
negative binomial	0.72 (0.59, 0.86)	4.5 (3, 9.5)	$r = 0.16$ (0.12, 0.20)	1675	0.14

the log-likelihood, given the shape parameter, is larger than its maximum minus 2 (also see the blue curve in figure 7).

Chi-square (goodness-of-fit) test. We bin the index cases with five to seven detectees, and all index cases with more than seven detectees to ensure that at least 10 observations are in one class.

We find that random mixing graphs and the Erdös–Renyi graph induce degree distributions with a rather light-weight tail. Therefore, these distributions do not fit the data appropriately. In the Poisson distribution, which is the degree distribution of an Erdösch–Renyi graph, the optimization routine even does not obtain a local maximum: it seems as in this case, the optimum is only assumed for an expectation $\mathbb{E}[K]$ that is unreasonably large or even tends to infinity.

The geometric distribution has a tail that is heavy enough to at least allow for a reasonable fit, but AIC as well as the chi-square (goodness of fit) test indicate that this model is rejected. The power-law (scale-free graph) is the first model that is at least weakly in line with the data: the tail of a power-law distribution may become heavy, and in this, there is a possibility to handle superspreading events appropriately. The AIC is worse but not too far from the winning model, and the p-value for the chi-square (goodness-of-fit) test is at least only in the range of 10^{-3}, and not less than 10^{-20}, as in the previous models.

The best model clearly is the negative binomial distribution, which is known to be an appropriate model for the number of contacts relevant to the transmission of respiratory infections [35]. The expected number of infectious contacts is small enough to be in the range of R_0, while—as expected—the over-dispersion is distinct. This model has the best AIC among all models, and the goodness-of-fit test does not reject this model.

It is interesting that the point estimate for p decreases if we choose models that have more mass in the tail. The reason is that the probability for k detectees scales with p^k, such that k small(er) needs to be balanced with more probability mass in the tail. A similar reason leads to smaller values for $\mathbb{E}[K]$ if the model distribution has more probability mass in the tail. However, this point estimate is rather similar for the power-law and the negative binomial and also does not heavily depend on the choice of R_0 (see 'sensitivity analysis'). In that, the range of p-estimate seems to be trustworthy. Moreover, the information in the data is sufficiently strong to point to a specific degree distribution (negative binomial), which was not clear from the beginning. As the data are rather simple, the information content could also have been too little to allow for distinct conclusions. That the negative binomial distribution, which is well known to be appropriate in this situation, is selected, is another sign that the estimates are trustworthy.

We first draw a contour plot (figure 7) indicating the point estimates (p and $\mathbb{E}[K]$) and also draw the cumulative empirical distribution vs. the cumulative theoretical distribution (figure 8), that is, on the x-axis, we plot the number of infectees, and on the y-axis the percentage of index cases which is the number of infected contact persons or less.

Also, these graphics clearly indicate that the power-law and the negative binomial distribution yield the best fit, where the negative binomial distribution is superior to the power-law.

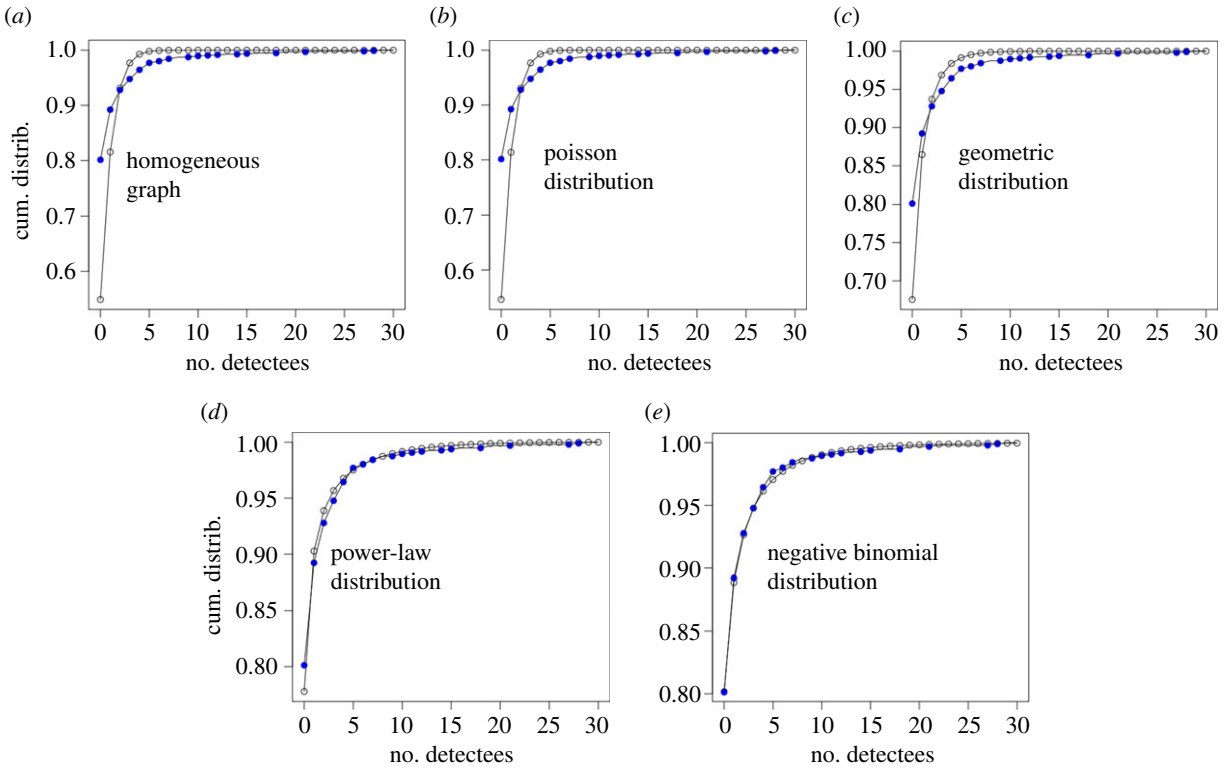

Figure 8. Cumulative empirical distribution (blue bullets) compared with the cumulative theoretical distribution (black circles). (*a*)–(*c*) Distributions for a homogeneous graph, a Poisson graph and a geometric distribution. (*d*,*e*) Distributions for a power-law distribution and a negative binomial distribution. Our proposed theory is indicated by the black circles.

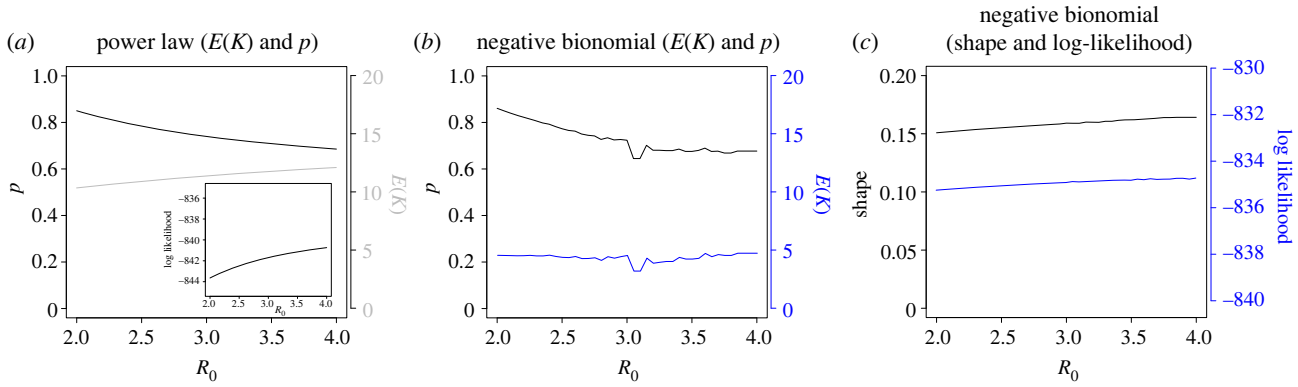

Figure 9. Sensitivity analysis illustrating changes in the point estimates (p and $E(K)$), and log-likelihood for different choices of R_0. (*a*) Power-law (with the log-likelihood as inlay), (*b*,*c*) negative binomial distribution.

6. Sensitivity analysis

We carried out a sensitivity analysis on how the estimated parameters depend on R_0. As illustrated in figure 9, the sensitivity analysis revealed that the estimated parameters in the studied models are not highly sensitive to the choice of R_0, at least for the power-law and negative distribution (the only degree distributions which meet the data satisfyingly). This observation is particularly noteworthy, as it underscores the robustness of these models in providing reliable estimates of epidemiological parameters, even when the initial assumptions about R_0 may vary. This characteristic is crucial in the context of real-world epidemiological studies, where the precise value of R_0 is often uncertain due to factors such as heterogeneous populations, changing contact patterns and varying degrees of intervention measures.

7. Discussion

In this paper, we present a graph-based method for estimating parameters in infectious disease models, offering valuable insights into the efficiency of contact tracing programmes and some information about local contact structures, and their implications for the spread of infectious diseases. By comparing various degree distributions and assessing their suitability for modelling disease spread, our analysis contributes to the ongoing development of improved parameter estimation techniques in graph-based models. Our findings complement and expand upon the work of Müller & Hösel [12] who estimated tracing probability in homogeneous populations by a maximum likelihood estimator and applied it to tuberculosis and chlamydia contact tracing data. In contrast to focusing on homogeneous random mixing populations, our work explores contact graphs in the form of trees which enables us to capture

the unique branching structure of infection transmission chains. This approach provides a more realistic representation of contact patterns at a microscopic level, allowing for a better understanding of the dynamics of infectious diseases and the effectiveness of contact tracing strategies.

Our comprehensive analysis of the COVID-19 contact tracing data from Karnataka, India, reveals that both scale-free network models and negative binomial distribution models offer a good fit for the data. The negative binomial distribution emerges as the most fitting model for the data, aligning with previous epidemiological research that has identified this distribution as a suitable assumption for the number of contacts relevant for the transmission of respiratory infections [35]. Furthermore, the observed over-dispersion in the number of secondary cases caused by individual index cases is accurately captured by the negative binomial distribution. This distribution is suitable for data where the variance exceeds the mean, reflective of scenarios where a small proportion of index cases are responsible for a disproportionate number of secondary infections. These findings resonate with the work of Gupta *et al.* [32] who had previously reported a clear over-dispersion in the data. Specifically, Gupta *et al.* [32] found that among 956 confirmed index cases, just 8.7% of cases, who had 14.4% of contacts, were responsible for 80% of all secondary cases. The power-law distribution also offers a reasonable fit, highlighting the potential relevance of scale-free networks in modelling infectious disease dynamics. In line with the principles of scale-free networks, our model highlights the role of a relatively small number of 'super-spreader' individuals, who have a significantly larger number of contacts and thus a higher likelihood of transmitting the infection to a larger pool of people. This also validates the findings of Gupta *et al.* [32] who suggested that super-spreaders may have played a more dominant role in the COVID-19 transmission in Karnataka, India.

Both the scale-free and negative binomial models allow for a thick or heavy tail, which is created by super-spreader events, and it is known in the case of airborne infection such as COVID-19 that super-spreader events and the over-dispersion of secondary cases have a significant impact on the effectiveness of contact tracing and surveillance schemes [31,32]. Regarding contact tracing, our models consistently indicate a high probability of successful tracing. These figures indeed raise questions regarding their realism and implications for epidemic dynamics. High probabilities for successful contact tracing suggest efficient public health measures in place, as well as the robustness of contact networks to facilitate tracing. However, a high tracing success probability which is not reflective of the high frequency of zero and few traced cases in the reported data might also raise concerns. In particular, regarding the choice of a tree-network where it is assumed that all infections present in the data are part of a single transmission tree. In real-world contact networks where individuals may be part of multiple overlapping transmission chains, contacts from outside the assumed tree are counted as part of it, potentially resulting in a higher number of traced contacts than what might occur in a non-tree-like network.

However, the high frequency of zero and few traced cases may have reflected the predominance of cases with a younger age-since-infection during the first wave due to certain preventive measures. Karnataka's contact tracing system was one of India's most effective, at least, during the early epidemic [36]. Considering the large proportion of close positive asymptomatic contacts at the time of testing, the low numbers present in the data could be indicative of effective pre-testing and preventive strategies at play. For instance, index cases identified and isolated quickly due to effective social distancing and lockdown measures, may result in fewer or no infectious secondary contacts, consequently leading to fewer or no traced cases. This may not necessarily reflect the efficacy of the contact tracing process itself but rather successful containment and prevention efforts that halted the spread from those index cases. This lack of distinction in the overall pre- and post-control strategies in our theory may have overestimated the tracing probability.

Furthermore, other epidemiological metric such as the reproduction number R_0 is critical to estimating the number of contacts to trace [37]. For instance, Hellewell *et al.* [38], who used simulations to study the feasibility of controlling COVID-19 outbreaks through the isolation of cases and contacts, found that to control 90% of outbreaks with a reproduction number of 2.5, 80% of contacts needed to be traced and isolated. Their research highlights the subtle role of reproduction number in contact tracing success, revealing that the probability of control increases at all levels of contact tracing when the reproduction number is reduced. Moreover, the Hellewell *et al.* [38] study emphasizes the significant impact of the number of initial cases on the likelihood of achieving control. Such insights highlight the complexities of contact tracing and its dependencies on various epidemiological and social factors. Additionally, while the high probabilities of contact tracing success implied by our model may raise questions about their realism, these figures are not unfounded. Our estimator also considers the interplay between the probability of tracing a contact once an index case is identified (p) and the probability of an infected individual being detected as an index case (p_{obs}). The latter has to be inherently lower, especially within close-knit contact networks such as family units.

In such networks, it is improbable for all contacts to become index cases; instead, tracing often occurs through one or a few known cases, underscoring why p_{obs} has to be substantially small. Given the nature of our methodology and the reported data, which concentrate on tracing only immediate contacts (one-step tracing), a higher p is plausible while maintaining lower p_{obs}. This is reflective of an efficient tracing system where immediate contacts are quickly identified, but not all are independently confirmed as index cases due to the close connection and simultaneous discovery through a single or limited number of initial cases. Despite p exceeding the ideal range for our approximations, the considerably smaller probability of any particular infected person becoming an index case ensures that our model's overall estimations are well suited. It is also crucial to understand that these probabilities, while informative, also underline the inherent complexities in predicting real-world outcomes. Variations in regional practices, public response, healthcare infrastructure and other socio-cultural factors play a significant role in the success of contact tracing endeavours. Therefore, while our model provides an essential tool for estimation, the results should be interpreted in conjunction with the broader epidemiological context and in light of other research findings for a holistic understanding.

Our modelling study and findings highlight the importance of selecting appropriate models for estimating tracing

probabilities and local contact structures in real-world scenarios. These estimates demonstrate the effectiveness of our graph-based method in capturing key epidemic parameters within heterogeneous and age-structured contact networks. The estimated degree ranges for the negative binomial and power-law distributions fall within plausible ranges found in the literature, supporting the validity of our approach in comparison with other studies that have examined contact tracing data [35,39,40]. Furthermore, sensitivity analysis on the estimated parameters with respect to R_0 for the power-law and negative binomial distribution shows limited sensitivity to the choice of R_0. This provides valuable insights into estimating key epidemiological and intervention parameters with greater confidence, enabling more effective public health strategies and interventions. All in all, our research contributes to the ongoing development of improved parameter estimation techniques in graph-based models for infectious disease dynamics. By using a graph-based approach and building upon the methods of previous studies, we have demonstrated the value of incorporating contact graph structures, such as trees, for a more accurate representation of contact patterns and infectious disease dynamics. The results of our analysis highlight the need to consider heterogeneity in individual-level contact networks when designing and evaluating contact tracing strategies.

Future research in this area could explore the incorporation of additional data sources and model refinements. A key area of refinement could be incorporating temporal dynamics or considering other types of contact graphs that better represent real-world contact patterns. The use of tree-like graphs in the current study, while mathematically convenient, is a simplification of reality. Contact patterns, particularly within clusters such as households or other social groups, often exhibit significant clustering and interconnectedness that a tree structure may fail to accurately capture. Our simulation study further highlights these concerns. For instance, the influence of triangles on the fraction of outside infections and the difference in the estimator's performance between Poisson and fixed degree models offer key insights. Such findings suggest a well-connected core group in certain models that leads to more outside infections, challenging our estimator's accuracy. Integrating well-suited graph structures with modelling techniques such as agent-based models or compartmental models, could provide a more comprehensive understanding of infectious disease dynamics and inform the design of more effective public health interventions.

Ethics. We have adhered to all ethical guidelines as laid out by the original study, which includes the appropriate ethical approvals for data collection and usage. Our research did not involve direct data collection from human subjects, and all analysis were conducted on anonymized, aggregated data without any personal identifiers to maintain individual confidentiality. The original dataset was collected in compliance with ethical standards of research, and our use of this dataset for secondary analysis is consistent with those standards.

Data accessibility. The analysis presented in this paper uses a publicly available dataset on COVID-19 contact tracing in Karnataka, India, originally reported by Gupta *et al.* [32]. The comprehensive description of the dataset, including the source of data, methodology for data compilation and ethical considerations, has been documented extensively by Gupta *et al.* [32]. The dataset can be accessed from the GitHub repository: https://github.com/CovidToday/covid19-karnataka.

Supplementary material is available online [41].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. A.O.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing—original draft, writing—review and editing; J.M.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing—review and editing; M.K.: formal analysis, funding acquisition, investigation, methodology, supervision, validation, visualization, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. The research was supported by a grant from the German Academic Exchange Service (DAAD) (A.O.) and by the Deutsche Forschungsgemeinschaft (DFG) through the TUM International Graduate School of Science and Engineering (IGSSE), GSC 81, as part of the project GENOMIE_QADOP (J.M.). M.K. acknowledges the support from the Horizon 2020 research and innovation funding programme (grant no. 101003480 (CORESMA)).

Acknowledgements. Portions of the work presented in this paper have previously been published in the PhD thesis authored by Augustine Okolie [42] and supervised by Johannes Müller. The thesis was submitted to the Technical University of Munich and is accessible online at https://mediatum.ub.tum.de/1661774. This current paper extends and enhances the findings discussed in the thesis, particularly those related to the likelihood estimator models and their application to both simulated data and empirical COVID-19 contact tracing data, as presented in chapters 3–5.

Appendix A. Full graph/random mixing

In the case of a full graph, we start with a fixed degree $K = N-1$, where N is the population size, and take the limit $N \to \infty$ and $\beta \to 0$ such that $R_0 = (N-1)\,\beta/(\mu+\sigma)$ is constant. Then,

$$(N-1)\,\hat{p}(a) = R_0\,\frac{\alpha+\sigma}{\alpha+\sigma-\beta}\left(\mathrm{e}^{-\beta a} - \mathrm{e}^{-(\alpha+\sigma)a}\right) \to R_0\,(1-\mathrm{e}^{-(\alpha+\sigma)\,a}).$$

Therefore, the binomial distribution approximates a Poisson distribution, and equation (4.2) becomes in the limit $N \to \infty$

$$P(T=i) = \int_0^\infty \mathrm{dpois}(i,\, R_0\,(1-\mathrm{e}^{-(\alpha+\sigma)\,a}))\; R_0\,(\alpha+\sigma)\,\mathrm{e}^{-R_0\,(\alpha+\sigma)a}\,\mathrm{d}a,$$

where $\mathrm{dpois}(i,\,\mu) = \mu^i\,\mathrm{e}^{-\mu}/i!$ is the probability function for the Poisson distribution.

Appendix B. Optimization process

Given a function $f(x)$, the aim of optimization is to find an x that either maximizes or minimizes $f(x)$. This process involves the computation of gradients and the Hessian matrix.

The gradient of a function is a vector that points in the direction of the greatest increase of that function. It is calculated as the vector of the first derivatives of the function with respect to each variable. The gradient of a function $f(x)$, where $x = [x_1, x_2, \ldots, x_n]$ is

$$\nabla f\,(x) = \left[\frac{\partial f}{\partial x_1},\, \frac{\partial f}{\partial x_2},\, \ldots,\, \frac{\partial f}{\partial x_n}\right]. \tag{B 1}$$

To ensure that the solution found is a local maximum and not a local minimum or a saddle point, the Hessian matrix is

used. The Hessian matrix is the square matrix of second-order partial derivatives of the function. Each element in the Hessian matrix is the second derivative of the function with respect to different variables. The Hessian matrix for the function $f(x)$ is

$$H(f(x)) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{B 2}$$

If the Hessian is positive definite (all eigenvalues are positive) at a point, then the function attains a local minimum at that point. If the Hessian is negative definite (all eigenvalues are negative), then the function attains a local maximum.

In the context of maximizing a likelihood function, we often convert the problem into a minimization problem by taking the negative of the likelihood function. This is due to the fact that many optimization algorithms are developed for minimization problems. The negative log-likelihood function becomes

$$-\mathcal{LL}(\boldsymbol{\mu} \mid i_\ell, \ell = 1, \ldots, n). \tag{B 3}$$

The goal now is to minimize this negative log-likelihood function, and the optimization problem becomes

$$\boldsymbol{\mu}^* = \underset{\boldsymbol{\mu}}{\operatorname{argmin}}\{-\mathcal{LL}(\boldsymbol{\mu} \mid i_\ell, \ell = 1, \ldots, n)\}. \tag{B 4}$$

The same principles of gradients and Hessians apply to this minimization problem. The gradient of the negative log-likelihood function should point in the direction of greatest decrease of the function. The Hessian, on the other hand, should be negative definite at the point of minimum.

In the case of the log-likelihood function, the optimization problem can be solved using iterative methods such as Newton's method or quasi-Newton methods, which make use of both the gradient and the Hessian of the function to find the minimum.

References

1. Khan MA. 2020 Parameter estimation and fractional derivatives of dengue transmission model. *AIMS Math.* **5**, 2758–2779. (doi:10.3934/math.2020178)
2. Little RJ *et al.* 2012 The prevention and treatment of missing data in clinical trials. *N. Engl. J. Med.* **367**, 1355–1360. (doi:10.1056/NEJMsr1203730)
3. Cantó B, Coll C, Sánchez E. 2009 Structural identifiability of a model of dialysis. *Math. Comput. Model.* **50**, 733–737. (doi:10.1016/j.mcm.2009.05.001)
4. Cantó B, Coll C, Sánchez E. 2011 Identifiability for a class of discretized linear partial differential algebraic equations. *Math. Prob. Eng.* **2011**, 249–250.
5. Craciun G, Pantea C. 2008 Identifiability of chemical reaction networks. *J. Math. Chem.* **44**, 244–259. (doi:10.1007/s10910-007-9307-x)
6. Blum MG, Tran VC. 2010 HIV with contact tracing: a case study in approximate Bayesian computation. *Biostatistics* **11**, 644–660. (doi:10.1093/biostatistics/kxq022)
7. O'Neill P, Roberts G and Bradford Univ.(United Kingdom), Dept. of Mathematics. 1997 *Bayesian inference for partially observed stochastic epidemics*. Bradford, UK: University of Bradford, School of Mathematical Sciences.
8. Götz T, Altmeier N, Bock W, Rockenfeller R, Wijaya KP. 2017 Modeling dengue data from Semarang, Indonesia. *Ecol. Complex.* **30**, 57–62. (doi:10.1016/j.ecocom.2016.12.010)
9. Agusto F, Khan M. 2018 Optimal control strategies for dengue transmission in Pakistan. *Math. Biosci.* **305**, 102–121. (doi:10.1016/j.mbs.2018.09.007)
10. Stollenwerk N, Aguiar M, Ballesteros S, Boto J, Kooi B, Mateus L. 2012 Dynamic noise, chaos and parameter estimation in population biology. *Interface Focus* **2**, 156–169. (doi:10.1098/rsfs.2011.0103)
11. Manou-Abi SM, Slaoui Y, Balicchi J. 2022 Estimation of some epidemiological parameters with the COVID-19 data of Mayotte. *Front. Appl. Math. Stat.* **8**, 67. (doi:10.3389/fams.2022.870080)
12. Müller J, Hösel V. 2007 Estimating the tracing probability from contact history at the onset of an epidemic. *Math. Popul. Stud.* **14**, 211–236. (doi:10.1080/08898480701612857)
13. Müller J, Kretzschmar M, Dietz K. 2000 Contact tracing in stochastic and deterministic epidemic models. *Math. Biosci.* **164**, 39–64. (doi:10.1016/S0025-5564(99)00061-9)
14. Dyson L, Marks M, Crook OM, Sokana O, Solomon AW, Bishop A, Mabey DC, Hollingsworth TD. 2018 Targeted treatment of yaws with household contact tracing: how much do we miss? *Am. J. Epidemiol.* **187**, 837–844. (doi:10.1093/aje/kwx305)
15. Tanaka T, Yamaguchi T, Sakamoto Y. 2020 Estimation of the percentages of undiagnosed patients of the novel coronavirus (SARS-CoV-2) infection in Hokkaido, Japan by using birth-death process with recursive full tracing. *PLoS ONE* **15**, e0241170. (doi:10.1371/journal.pone.0241170)
16. Okolie A, Müller J. 2020 Exact and approximate formulas for contact tracing on random trees. *Math. Biosci.* **321**, 108320. (doi:10.1016/j.mbs.2020.108320)
17. Bertacchini F, Bilotta E, Pantano PS. 2020 On the temporal spreading of the SARS-CoV-2. *PLoS ONE* **15**, e0240777. (doi:10.1371/journal.pone.0240777)
18. Chondros C, Nikolopoulos SD, Polenakis I. 2022 An integrated simulation framework for the prevention and mitigation of pandemics caused by airborne pathogens. *Netw. Model. Anal. Health Inform. Bioinform.* **11**, 42. (doi:10.1007/s13721-022-00385-z)
19. Cuevas-Maraver J, Kevrekidis PG, Chen QY, Kevrekidis GA, Villalobos-Daniel V, Rapti Z, Drossinos Y. 2021 Lockdown measures and their impact on single- and two-age-structured epidemic model for the COVID-19 outbreak in Mexico. *Math. Biosci.* **336**, 108590. (doi:10.1016/j.mbs.2021.108590)
20. Kovacevic R, Stilianakis NI, Veliov VM. 2022 A distributed optimal control epidemiological model applied to COVID-19 pandemic. *SIAM J. Optim.* **60**, S221–S245. (doi:10.1137/20M1373840)
21. Modi K, Umate L, Makade K, Dubey RS, Agarwal P. 2021 Simulation based study for estimation of COVID-19 spread in India using SEIR model. *J. Interdiscipl. Math.* **24**, 245–258. (doi:10.1080/09720502.2020.1838059)
22. Kevrekidis PG, Cuevas-Maraver J, Drossinos Y, Rapti Z, Kevrekidis GA. 2021 Reaction-diffusion spatial modeling of COVID-19: Greece and Andalusia as case examples. *Phys. Rev. E* **104**, 024412. (doi:10.1103/PhysRevE.104.024412)
23. Green DM, Kiss IZ. 2010 Large-scale properties of clustered networks: implications for disease dynamics. *J. Biol. Dyn.* **4**, 431–445. (doi:10.1080/17513758.2010.487158)
24. Keeling MJ. 1999 The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. Lond. B* **266**, 859–867. (doi:10.1098/rspb.1999.0716)
25. Kiss IZ, Miller JC, Simon PL. 2017 *Mathematics of epidemics on networks*, vol. 598. Cham, Switzerland: Springer, p. 31.
26. Newman ME. 2002 Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 016128. (doi:10.1103/PhysRevE.66.016128)
27. Karrer B, Newman ME. 2010 Message passing approach for general epidemic models. *Phys. Rev. E* **82**, 016101. (doi:10.1103/PhysRevE.82.016101)
28. Feld SL. 1991 Why your friends have more friends than you do. *Am. J. Soc.* **96**, 1464–1477. (doi:10.1086/229693)
29. Brauer F, Van den Driessche P, Wu J. 2008 *Mathematical epidemiology*, vol. 1945. Berlin, Germany: Springer-Verlag.

30. Christakis NA, Fowler JH. 2010 Social network sensors for early detection of contagious outbreaks. *PLoS ONE* **5**, e12948. (doi:10.1371/journal.pone.0012948)
31. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. 2005 Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359. (doi:10.1038/nature04153)
32. Gupta M *et al.* 2022 Contact tracing of COVID-19 in Karnataka, India: superspreading and determinants of infectiousness and symptomatic infection. *PLoS ONE* **17**, e0270789. (doi:10.1371/journal.pone.0270789)
33. Bell BP *et al.* 1994 A multistate outbreak of *Escherichia coli* O157: H7–associated bloody diarrhea and hemolytic uremic syndrome from hamburgers: the Washington experience. *J. Am. Med. Assoc.* **272**, 1349–1353. (doi:10.1001/jama.1994.03520170059036)
34. Locatelli I, Trächsel B, Rousson V. 2021 Estimating the basic reproduction number for COVID-19 in Western Europe. *PLoS ONE* **16**, e0248731. (doi:10.1371/journal.pone.0248731)
35. Mossong J *et al.* 2008 Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**, e74. (doi:10.1371/journal.pmed.0050074)
36. Priya A *et al.* 2020 Laboratory surveillance for SARS-CoV-2 in India: performance of testing & descriptive epidemiology of detected COVID-19, January 22 - April 30, 2020. *Indian J. Med. Res.* **151**, 424. (doi:10.4103/ijmr.IJMR_1896_20)
37. Linka K, Peirlinck M, Kuhl E. 2020 The reproduction number of COVID-19 and its correlation with public health interventions. *Comput. Mech.* **66**, 1035–1050. (doi:10.1007/s00466-020-01880-8)
38. Hellewell J *et al.* 2020 Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Global Health* **8**, e488–e496. (doi:10.1016/S2214-109X(20)30074-7)
39. Soetens L, Klinkenberg D, Swaan C, Hahné S, Wallinga J. 2018 Real-time estimation of epidemiologic parameters from contact tracing data during an emerging infectious disease outbreak. *Epidemiology* **29**, 230–236. (doi:10.1097/EDE.0000000000000776)
40. Fyles M *et al.* 2021 Using a household-structured branching process to analyse contact tracing in the SARS-CoV-2 pandemic. *Phil. Trans. R. Soc. B* **376**, 20200267. (doi:10.1098/rstb.2020.0267)
41. Okolie A, Müller J, Kretzschmar M. 2023 Parameter estimation for contact tracing in graph-based models. Figshare. (doi:10.6084/m9.figshare.c.6927385)
42. Okolie AO. 2022 Contact tracing on stochastic graphs. PhD thesis, Universitätsbibliothek der TU München, Germany.