Bioinformatics Advances, 2023, **00**, vbad082
https://doi.org/10.1093/bioadv/vbad082
Advance access publication 3 July 2023
Original Article

OXFORD

Disease bioinformatics/Translational medicine

Network-augmented compartmental models to track asymptomatic disease spread

Devavrat Vivek Dabke [1,†], Kritkorn Karntikoon [2,†], Chaitanya Aluru [2], Mona Singh [2] and Bernard Chazelle [2,*]

[1]The Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA
[2]Department of Computer Science, Princeton University, Princeton, NJ 08544, USA

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

*To whom correspondence should be addressed.

Associate Editor: Thomas Lengauer

Abstract

Summary: A major challenge in understanding the spread of certain newly emerging viruses is the presence of asymptomatic cases. Their prevalence is hard to measure in the absence of testing tools, and yet the information is critical for tracking disease spread and shaping public health policies. Here, we introduce a framework that combines classic compartmental models with travel networks and we use it to estimate asymptomatic rates. Our platform, traSIR ("tracer"), is an augmented susceptible-infectious-recovered (SIR) model that incorporates multiple locations and the flow of people between them; it has a compartment model for each location and estimates of commuting traffic between compartments. TraSIR models both asymptomatic and symptomatic infections, as well as the dampening effect symptomatic infections have on traffic between locations. We derive analytical formulae to express the asymptomatic rate as a function of other key model parameters. Next, we use simulations to show that empirical data fitting yields excellent agreement with actual asymptomatic rates using only information about the number of symptomatic infections over time and compartments. Finally, we apply our model to COVID-19 data consisting of reported daily infections in the New York metropolitan area and estimate asymptomatic rates of COVID-19 to be ∼34%, which is within the 30–40% interval derived from widespread testing. Overall, our work demonstrates that traSIR is a powerful approach to express viral propagation dynamics over geographical networks and estimate key parameters relevant to virus transmission.

Availability and implementation: No public repository.

Contact: chazelle@cs.princeton.edu

1 Introduction

At the outset of the COVID-19 pandemic, the prevalence of asymptomatic cases among infections was estimated to lie anywhere between 17% and 81% (Nogrady, 2020). Given the importance of this parameter for early health policy decisions (Nishiura *et al.*, 2020), such a high level of uncertainty was a major roadblock. With testing now widely available, this issue has largely dissipated, with estimates of asymptomatic rates between ∼30% and ∼40% (Ma *et al.*, 2021; Shang *et al.*, 2022). To prevent such difficulties in future epidemics, it would be highly beneficial to have computational tools for estimating the asymptomatic rate of infected individuals right at the beginning of an epidemic.

Infectious disease spread is classically modeled using compartmental models. The population is assigned to distinct compartments (e.g. the susceptible, infectious and recovered compartments in the widely studied SIR model) (Kermack and McKendrick, 1927), with rates at which individuals move from one compartment to another. When applying these compartment-based epidemiological models, it is impossible to predict the true prevalence of a virus early on in a pandemic without widespread random testing: indeed, even a tiny fraction of a population showing symptoms for the disease is compatible with a widespread infection. To estimate via computational modeling the fraction of infectious individuals that are asymptomatic, or the *asymptomatic rate* ρ, requires additional information. Here, we show that considering information about how a virus spreads in a spatial manner—not just between compartments at a single location—can be leveraged to estimate ρ. The intuition is that, while individuals travel between locations and this contributes to viral spread, individuals who feel sick (i.e. are symptomatic) tend to curb travel, which in turn yields a distinguishing observable between symptomatic and asymptomatic carriers.

In this article, we introduce traSIR (pronounced "tracer"), a network traffic-based SIR model, which combines the classic SIR compartmental model with network modeling. In traSIR, we have a network where each node is a location (e.g. a county or ZIP Code), each location is associated with a compartmental model and edges in the network represent frequent travel between the locations (e.g. commuting). TraSIR additionally models asymptomatic and symptomatic infections, together with a dampening effect on viral spread for symptomatic infections. Our primary contribution is to demonstrate the utility of traSIR in

Received: May 20, 2023. **Revised:** June 20, 2023. **Editorial Decision:** June 26, 2023. **Accepted:** July 1, 2023

estimating the asymptomatic rate of an infectious disease using only knowledge about symptomatic infections across geographic locations, as well as information about typical travel between locations.

We begin with theoretical results relating the asymptomatic rate of infection to other key parameters of the model (e.g. infection and recovery rates). Since these key parameters are not known *a priori* and must be estimated from the data, we next assess how well parameters of a traSIR model can be estimated using only knowledge about symptomatic infections. In particular, we simulate disease spread using traSIR, and then perform empirical parameter estimation using the number of symptomatic infections over time across locations to estimate the asymptomatic rate ρ. Across a wide range of parameters, we find excellent agreement between the actual and estimated ρ values. Finally, we analyze the number of reported COVID-19 infections across the New York metropolitan area during the first wave, from March 1, 2020 to September 17, 2020.

The method behind traSIR seeks to combine topological flow information with diagnostic data and behavioral variations. It makes use of a number of observable nonlinearities: (i) in the absence of public health measures, a multiplicative decrease in the symptomatic rate causes a forward time-shift in the infection curve relative to its measurable baseline; (ii) detection of carriers grow superlinearly in the number of symptomatic cases; (iii) the number of newly symptomatic cases is largely determined by the asymptomatic neighbors in the network and (iv) asymptomatic carriers have a different transmissibility rate (Li *et al.*, 2020). Our platform, traSIR, is the first of its kind to integrate county-level data with a commuter network on a large scale to recover critical epidemiological characteristics directly from network dynamics, in particular the asymptomatic rate.

1.1 Further background

Standard epidemiological models have previously been extended to account for disease spread across space, but the medium has typically been assumed to be homogeneous (where the population is treated as one large group, as opposed to interacting subpopulations) (Brauer *et al.*, 2008; Eletreby *et al.*, 2020), leading to a diffusive process. Typically, the speed of a wave across the population grows in proportion to the square root of the reproduction number and the diffusion coefficient. Epidemics have also previously been studied in random graphs and scale-free networks (Ajelli *et al.*, 2010; Barrat *et al.*, 2008). Previous work has also considered the correlation of viral spread with changing commuting patterns as well as signals from social media or search engines (Byambasuren, 2020; Li *et al.*, 2020; Poletti *et al.*, 2021; Subramanian and Pascual, 2021; Sun *et al.*, 2021; Zhan *et al.*, 2018); other approaches have integrated network effects into compartmental models (Ameri and Cooper, 2019; Barabási, 2013; Ding *et al.*, 2021; Dottori and Fabricius, 2015; Liu *et al.*, 2018). However, they lack any symmetry-breaking mechanism for distinguishing between symptomatic and asymptomatic carriers. This is precisely what traSIR offers.

Even agent-based modeling does not directly resolve this issue. While agent-based modeling certainly provides enhanced resolution and a distribution of outcomes (certain aspects of which we leverage by decomposing our model to the county level), agent-based modeling has not been able to properly estimate asymptomatic spread (Kerr *et al.*, 2021). One significant downside of agent-based modeling is that it is also computationally expensive, even when leveraging vectorized features to control spread mechanisms (Dabke and Arroyo, 2016).

1.2 Symmetry-breaking and asymptomatic spread

In order to properly estimate the asymptomatic rate, we need to be able to distinguish between symptomatic and asymptomatic spread. The model we present in this article allows us to distinguish between these two types of spread via different interaction patterns: we can have varying commuting and travel patterns among asymptomatic and symptomatic populations.

Our results show that under simple assumptions or static parameters, we can still distinguish the spread within these two populations. This implies that having some symmetry-breaking is paramount and while we have ample room for refinement, mere existence of symmetry-breaking is sufficient for estimating the asymptomatic rate accurately.

The asymptomatic rate can be estimated clinically (Johansson *et al.*, 2021; Martinelli *et al.*, 2022; Sah *et al.*, 2021); while this is the most accurate approach, it requires large-scale surveillance testing, which can be prohibitively resource-intensive. In contrast, computational approaches tend to add compartments to traditional SIR models, which is also our strategy. In Li *et al.* (2021), they add multiple compartments; we add just one asymptomatic compartment, which simplifies our model and increases computational tractability. Our approach is most similar to Layton and Sadria (2022), but we generalize to a larger geography and incorporate commuter data. For any modeling task, especially with a novel infection, aggregating different modeling approaches and data sources usually yields a strong consensus estimate; therefore, we hope to add to the literature by providing a modeling refinement and an additional estimate of the asymptomatic rate of spread to the ensemble.

2 Methods

2.1 The model

We show how to embed the classic SIR epidemiological model (Kermack and McKendrick, 1927) within a geographic network with known travel rates. The network $G = (V, E)$ is a directed graph joining N nodes (typically, counties), whose edges are annotated with the corresponding mean traffic rates of commuters. The edge set E includes all the pairs (i, j) such that residents of county i commute to work in county j. We assume the availability of an N-by-N stochastic "commute" matrix M, such that M_{ij} indicates the probability that someone commutes from county i to county j on a typical workday.

On day t, we denote the number of susceptible and recovered individuals in county i by $s_i(t)$ and $r_i(t)$, respectively. Among the $f_i(t)$ carriers of the virus in the county, we distinguish between the $c_i(t)$ of them who show symptoms and the $a_i(t) = f_i(t) - c_i(t)$ who do not. The population size in county i is denoted by $n_i = f_i(t) + s_i(t) + r_i(t)$ and is assumed fixed over the period under investigation. For convenience, we may write the right-hand side as $\sum_{x\in\{f,s,r\}} x_i(t)$.

The commute matrix M is blind to the health status of commuters. Symptomatic people tend to travel less, however, and this change has great effect on contagion. To capture

this phenomenon, we introduce the *decommute rate* $\delta \in [0,1]$ as a measure of the propensity of people feeling sick to stay home:

$$M^c = (1-\delta)M + \delta\mathbb{I}. \tag{1}$$

where $\mathbb{I}$ represents the identity matrix. Note that, if $\delta = 0$, being symptomatic has no bearing on commuting. The matrix M^c is a symmetry-breaking device which allows to distinguish between sick virus carriers and the rest. This difference creates observable nonlinearities in the viral dynamics that we can exploit to estimate the asymptomatic rate ρ. While our model could allow M^x to be a function of time, we simply require M^c to be different from the other transition matrices in order to get the benefit of symmetry-breaking. Therefore, we let all of them be static and set $M^s = M^a = M^r = M$. In the results reported here, we set δ to $8/9$.

2.1.1 The chronology of infection

Instead of stating the model all at once, we introduce it one piece at a time, following its natural chronology. We fix a county i and trace the changes in the main state variables s, c, a, r, f, n. We use specific times for illustrative purposes only.

- **Step 1:** At 8 am on day t, all commuters are ready to go to work. We have $f_i(t) = c_i(t) + a_i(t)$ and $\sum_{x\in\{s,c,a,r\}} x_i(t) = n_i(t) = n_i$.
- **Step 2:** At 9 am, commuters are at work. This changes the local population into a transient one, which we denote with a "hat." By definition of the commute matrix, $\hat{x}_i(t) = \sum_j M^x_{ji} x_j(t)$ for $x = s, c, a, r$, with $\hat{f}_i(t) = \sum_{x\in\{c,a\}} \hat{x}_j(t)$ and $\hat{n}_i(t) = \sum_{x\in\{s,f,r\}} \hat{x}_j(t)$. The transient population at county i will now get to mix all day at work and spread the infection among itself.
- **Step 3:** At 5 pm, commuters go home. The new population at county i is denoted with a "bar." It consists of the same n_i people present at 8 am, but with a different health status distribution. Take the set of infected individuals: it includes the $f_i(t)$ carriers from 8 am plus the newly infected. The latter consist of the subset of the $s_i(t)$ susceptible individuals who caught the virus by commuting to county j and got exposed to a carrier in the transient population of j. Note that this includes the case $j = i$ of non-commuters who were exposed to infected visitors. The chance of anyone getting sick in this fashion is $\varphi_j(t) := \beta\hat{f}_j(t)/\hat{n}_j(t)$, where $0 < \beta < 1$ measures the transmission rate: it is the average number of contacts per person per day times the probability of transmission in a contact between an infected person and a susceptible one. (The model can easily accommodate a time-varying rate β. The reason for keeping it fixed is to decouple the baseline socialization rate from its pandemic-induced variations via the commute matrices.)

  The number of newly infected residents of county i is the sum, over all j, of the number of commuters from county i who went to county j and got infected there: therefore, it is equal to $s_i(t)\psi_i(t)$, where $\psi_i(t) := \sum_j M_{ij}\varphi_j(t) < 1$ denotes the *worktime infectivity rate*: it is the probability that a commuter from i catches the virus at work.

  We have $\overline{f}_i(t) = f_i(t) + s_i(t)\psi_i(t)$. Since a fraction ρ of these new infections are asymptomatic, we have

$$\begin{cases} \overline{s}_i(t) = s_i(t)(1-\psi_i(t)) \\ \overline{c}_i(t) = c_i(t) + (1-\rho)s_i(t)\psi_i(t) \\ \overline{a}_i(t) = a_i(t) + \rho s_i(t)\psi_i(t). \end{cases} \tag{2}$$

- **Step 4:** At 8 am on day $t+1$, further mixing will have occurred in county i since the previous evening. A fraction γ of the infected people will have recovered by then. Writing

$$u_i(t) = \beta\overline{s}_i(t)\left(\frac{\overline{c}_i(t) + \overline{a}_i(t)}{n_i}\right),$$

we have

$$\begin{cases} s_i(t+1) = \overline{s}_i(t) - u_i(t) \\ c_i(t+1) = (1-\gamma)\overline{c}_i(t) + (1-\rho)u_i(t) \\ a_i(t+1) = (1-\gamma)\overline{a}_i(t) + \rho u_i(t) \\ r_i(t+1) = r_i(t) + \gamma\overline{c}_i(t) + \gamma\overline{a}(t). \end{cases} \tag{3}$$

We note that traSIR involves two rounds of mixing: the first one in the daytime accounts for intercounty infection (via commuting); the second one (nighttime) models intracounty infection (within each county). For simplicity, we model recovery in the latter only. (For this reason, our value of γ might differ from the standard one by a factor of 2.)

2.1.2 traSIR in vector form

We can give a compact description of the model. Let $x(t)$ denote the row vector with N coordinates $x_i(t)$, for $x \in \{s, c, a, r, n\}$. We define the row vectors

$$\begin{cases} \hat{x}(t) = x(t)M^x(t), \text{ for } x \in \{s,c,a,r\} \\ \hat{f}(t) = \sum_{x\in\{c,a\}} \hat{x}(t)\,; \;\; \hat{n}(t) = \sum_{x\in\{s,f,r\}} \hat{x}(t) \\ \varphi(t) = \beta\hat{f}(t) \oslash \hat{n}(t)\,; \;\; \psi_t = \varphi(t)M^s(t)^T. \end{cases} \tag{4}$$

Using the symbols $\otimes$ and $\oslash$ to refer to component-wise vector multiplication and division, respectively, we have

$$\begin{cases} \overline{s}(t) = s(t) - s(t) \otimes \psi(t) \\ \overline{c}(t) = c(t) + (1-\rho)s(t) \otimes \psi(t) \\ \overline{a}(t) = a(t) + \rho s(t) \otimes \psi(t). \end{cases} \tag{5}$$

For $u(t) := \beta\overline{s}(t) \otimes (\overline{c}(t) + \overline{a}(t)) \oslash (n_1, \ldots, n_N)$,

$$\begin{cases} s(t+1) = \overline{s}(t) - u(t) \\ c(t+1) = (1-\gamma)\overline{c}(t) + (1-\rho)u(t) \\ a(t+1) = (1-\gamma)\overline{a}(t) + \rho u(t) \\ r(t+1) = r(t) + \gamma\overline{c}(t) + \gamma\overline{a}(t). \end{cases} \tag{6}$$

2.2 Parameter estimation

Given a commute network and daily symptomatic infections across each node in the network, we develop an approach for estimating the asymptomatic rate ρ. The estimation algorithm can be viewed as a two-player game in which participants take turns updating their current estimates of (β, γ) and ρ, respectively. Recall that β, γ measure the infection and recovery rate, respectively. We assume that all counties have the same value of β and γ. The updating is driven by grid search (and gradient descent) with respect to a normalized mean-square

Algorithm 1

procedure ESTIMATE(c)

$\rho^* \leftarrow 0.5$
for $\ell = 1, 2, \ldots, j_{\max}$ **do**
▷ use grid search to optimize (β^*, γ^*) via normalized mean-square loss function at initial county i_0
$(\beta^*, \gamma^*) \leftarrow \text{argmin}_{(\beta,\gamma)} \mathcal{L}(c_{i_0}, \hat{c}_{i_0}(\beta, \gamma, \rho^*))$
▷ use grid search to optimize ρ^* via normalized mean-square loss function across all counties
$\rho^* \leftarrow \text{argmin}_\rho \sum_{i=1}^N \mathcal{L}(c_i, \hat{c}_i(\beta^*, \gamma^*, \rho))$
▷ gradient descent on ρ^*
$g(x) := \sum_{i=1}^N \mathcal{L}(c_i, \hat{c}_i(\beta^*, \gamma^*, x))$
$k \leftarrow 0$; $\tau \leftarrow \infty$
while $\tau > \tau_{\min}$ & $k < k_{\max}$ **do**
$\tau \leftarrow \varepsilon(dg/dx)(\rho^*)$
$\rho^* \leftarrow \max\{0, \rho^* - \tau\}$
$k \leftarrow k + 1$
return $(\beta^*, \gamma^*, \rho^*)$

loss function, which is computed for a node k across all time points as follows:

$$\mathcal{L}(c, \hat{c}) = \sum_{t=1}^{T} \left(\frac{c(t)}{||c||_\infty} - \frac{\hat{c}(t)}{||\hat{c}||_\infty} \right)^2, \tag{7}$$

where c is the vector in $[0,1]^T$ whose coordinate $c(t)$ denotes the recorded rate of symptomatic cases in the population in some given county k at time t. We write $c = c_k$ when disambiguation is needed.

The normalization makes the loss invariant under scaling. This is a necessary feature given the noise in the data. Of highest concern is the corruption of the official figures caused by the inclusion of reported asymptomatic cases via testing and the exclusion of symptomatic patients who do not seek a diagnosis. We assume that the signal-to-noise ratio remains constant over time; hence, that the time series c is available up to an unknown scaling factor. The normalization factors out that uncertainty.

The vector $\hat{c} = \hat{c}(\beta, \gamma, \rho)$ is the traSIR-predicted counterpart to the factual vector c; the matrix M and the decommute rate, defined in (1), are fixed. We assume that the infection is seeded at county i_0. With ρ expected to exert a relatively minor influence on the transmission/recovery parameters at the seeded node, it is natural to base the estimate of (β, γ) on the time series c_{i_0}.Within Algorithm 1, we set $j_{\max} = 3$ (convergence is quick). The grid search is over a discrete space of size 10^3 for ρ and 10^4 for (β, γ). The number of gradient descent steps is $k_{\max} = 10^3$; the gradient descent threshold is $\tau_{\min} = 10^{-12}$ and the learning rate is $\varepsilon = 10^{-4}/NT$. Note that the output $(\beta^*, \gamma^*, \rho^*)$ will be referred to as the estimated parameters $(\hat{\beta}, \hat{\gamma}, \hat{\rho})$ in the text.

2.3 Actual data

For the commute network and population data, we rely on the most recent (pre-COVID) *American Community Survey* from the U.S. Census Bureau (American Community Survey [ACS], 2015a,b). The nodes in the network represent the counties; the edges are directed and weighted in proportion to the number of residents who live in the source county and work in the destination county. We clean up the data by removing all the edges associated with fewer than 10 000 commuters. From the resulting graph, we extract the largest weakly connected component, which in this case corresponds to the New York Metropolitan Area. It consists of 44 counties: a visualization of which can be seen in Figure 1. For the infection data, we use the New York Times COVID-19 tracker and focus on the 200 days between March 1, 2020 and September 17, 2020 (Connell, 2022; The COVID Tracking Project at The Atlantic, 2022; The New York Times, 2022) (https://www.census.gov/data/tables/2015/demo/metro-micro/commuting-flows-2015.html and https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2015/5-year.html).

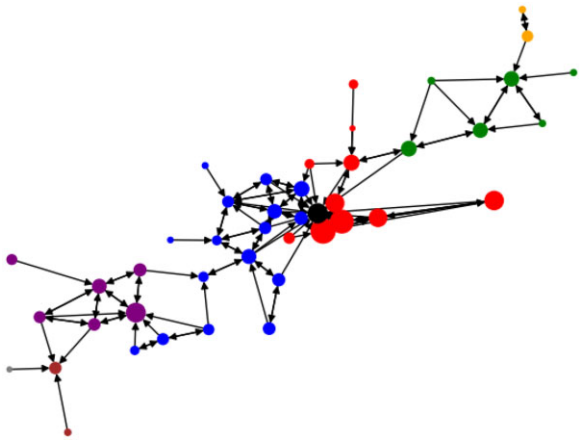

Figure 1. New York City metropolitan area: Each state is colored differently, with, at the center, Manhattan in black. The node size corresponds to the population in that county. The nodes are positioned according to the geographic center of each county

2.4 Simulated data

We generate 481 low-discrepancy values of β, γ and ρ, where $0.2 \leq \beta, \rho \leq 0.8, 0.01 \leq \gamma \leq 0.7$, using Sobol sequences from the SciPy package (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.qmc.Sobol.html). For each of the 481 combinations of parameters, we run traSIR with the corresponding parameters for 150 timesteps on the New York Metropolitan area population and network data, assuming that there is a single infected individual in New York County (Manhattan). We further corrupt the resulting symptomatic population sizes by a fixed scalar that is unknown to the algorithm.

For validation, we run Algorithm 1 on the corrupted simulation to produce the estimated parameters $(\hat{\beta}, \hat{\gamma}, \hat{\rho})$. We evaluate the accuracy and tabulate the residuals between the estimation and actual parameters (β, γ, ρ).

3 Results

The main contribution of this article is to demonstrate empirically that a network-based epidemiological model can uncover key parameters of a contagious disease. We provided an intuitive explanation for why this might be possible as long as a symmetry-breaking mechanism is in place for distinguishing among different types of virus carriers. Before we discuss the

empirical evidence and validate our approach, we provide a succinct mathematical foundation for our claim.

3.1 Theoretical analysis

We fix the county i and the time t and we drop all mention of t when it is understood from the context. By Equations (2) and (3),

$$\begin{aligned} f_i(t+1) &= (1-\gamma+\beta\overline{s}_i/n_i)\overline{f}_i \\ &= (1-\gamma+\beta s_i(1-\psi_i)/n_i)(f_i+s_i\psi_i) \\ &= (1-\gamma+\beta s_i/n_i)f_i \\ &\quad +(1-\gamma+\beta(s_i-f_i)/n_i)s_i\psi_i-(\beta/n_i)(s_i\psi_i)^2, \end{aligned} \tag{8}$$

where

$$\begin{aligned} \psi_i &= \sum_j M^s_{ij}\varphi_j(t) \\ &= \beta\sum_j M_{ij}\frac{\hat{f}_j}{\hat{n}_j} \\ &= \beta\sum_j M_{ij}\frac{\sum_k(f_k-\delta c_k)M_{kj}+\delta c_j}{\sum_k(n_k-\delta c_k)M_{kj}+\delta c_j}. \end{aligned}$$

Recall that $\hat{f}_i(t)$ denotes the number of infected individuals in the transient population at county i at the end of the morning commute. Let $f'_i=\sum_k f_k M_{ki}$ be the number it would have been if we had $\delta=0$ and hence $M^c=M$; we derive $n'_i=\sum_k n_k M_{ki}$ from $\hat{n}_i(t)$ likewise. We have

$$\begin{cases} \hat{f}_j(t)=\sum_k(f_k-\delta c_k)M_{kj}+\delta c_j=f'_j-\delta(1-\rho)g_j \\ \hat{n}_j(t)=\sum_k(n_k-\delta c_k)M_{kj}+\delta c_j=n'_j-\delta(1-\rho)g_j, \end{cases} \tag{9}$$

where $g_j=f'_j-f_j$. This allows us to rewrite ψ_i as

$$\psi_i=\beta\sum_j M_{ij}\left(\frac{f'_j-\delta(1-\rho)g_j}{n'_j-\delta(1-\rho)g_j}\right). \tag{10}$$

The worktime infectivity rate ψ_i plays a key role in traSIR. If $M=\mathbb{I}$, then $\psi_i=\beta f_i/n_i$ is the usual infectivity rate in the classic SIR model. Take the case of an arbitrary matrix M and set $\delta=0$. Denote by $\mathbf{E}_{j(i)}$ the expectation operator indexed by i and defined by M_{ij}, for $j=1,\ldots,N$. Likewise, we introduce the expectation operator $\mathbf{E}_{k(j)}$, indexed by j and defined by $n_k M_{kj}/\sum_l n_l M_{lj}$, for $k=1,\ldots,N$. It follows that

$$\begin{aligned} \psi_i^{|\delta=0} &= \beta\sum_j M_{ij}\sum_k\left(\frac{n_k M_{kj}}{\sum_l n_l M_{lj}}\right)\frac{f_k}{n_k} \\ &= \beta\mathbf{E}_j\mathbf{E}_{k(j)}\frac{f_k}{n_k}. \end{aligned} \tag{11}$$

We conclude that, when decommuting is withheld ($\delta=0$), ψ_i is an average of infection ratios f_k/n_k over counties adjacent to i or adjacent to the latter. This two-degree of separation corresponds to individuals from distinct counties meeting at work in a third county. The same idea holds for $\delta>0$, but with corrective terms that we discuss below.

In epidemiology, an important characteristic of an infection is the basic reproduction number R_0, which measures the average number of cases generated by an infected individual (Anderson and May, 1991). At the outset of the pandemic, we can use $f_i(t+1)/f_i(t)$ as a proxy for the reproduction number R_0 associated with county i. In a classical SIR model, R_0 has the form β/γ, but in TraSIR, it follows from (8) that

$$\begin{aligned} R_0 &= 1-\gamma+\frac{\beta s_i}{n_i}+\left(\frac{1-\gamma+\beta(s_i-f_i)/n_i}{f_i}\right)s_i\psi_i \\ &\quad -\frac{\beta}{f_i n_i}(s_i\psi_i)^2. \end{aligned} \tag{12}$$

Together, Equations (10) and (12) form a system $\mathcal{S}(\rho)=0$, which in theory allows us to recover the asymptomatic rate ρ from β, γ and R_0. It is noteworthy that this requires decommuting. The system $\mathcal{S}$ cannot be solved for ρ in closed form. Using traSIR for estimation can thus be viewed as a numerical solver for $\mathcal{S}$.

3.2 Simulations

We demonstrate that Algorithm 1 can accurately recover the infection rate β, recovery rate γ and asymptomatic rate ρ in simulated infections across a wide range of parameters, using just knowledge about the network and the numbers of symptomatic infected individuals. For each of simulations resulting from many combinations of parameters (see Methods), we will use the number of symptomatic individuals for each county over time. In practice, the actual number of symptomatic individuals is larger than the number reported, we multiply each of the resulting symptomatic population sizes by a fixed scalar (unknown to the algorithm), and then run Algorithm 1 to produce the estimated parameters $(\hat{\beta},\hat{\gamma},\hat{\rho})$.

We find excellent agreement between the actual parameters β, γ and ρ and their estimates $(\hat{\beta},\hat{\gamma},\hat{\rho})$ (Fig. 2). Figure 2 shows a scatter plot of an estimated parameter against the corresponding synthetic parameter for the New York area. The Pearson correlation coefficient is 0.9996 between β and its predicted value. For γ and ρ, it is 0.9983 and 0.9915, respectively. The absolute residual across all starting parameters has mean 0.023 and standard deviation 0.017. The absolute residual mean and standard deviation for β are 0.0032 and 0.00246; for γ are 0.0065 and 0.0064 and for ρ are 0.0158 and 0.0163.

3.3 Applications to COVID-19 data

Having validated our estimation technique on simulated data, we now apply Algorithm 1 to daily infection numbers from the New York Metropolitan area (see Methods), and estimate the asymptomatic rate ρ, a parameter of critical importance to health policymakers. We find:

$$\beta=0.320 \quad ; \quad \gamma=0.046 \quad ; \quad \rho=0.345\,.$$

Using these parameters, we also compared the traSIR-simulated symptomatic infection count with the real reported infection count across the New York metropolitan area (Fig. 3), and find good agreement.

4 Discussion

We have shown via theoretical analysis and simulation that a network-augmented compartmental model can effectively estimate the asymptomatic rate of viral infections using only

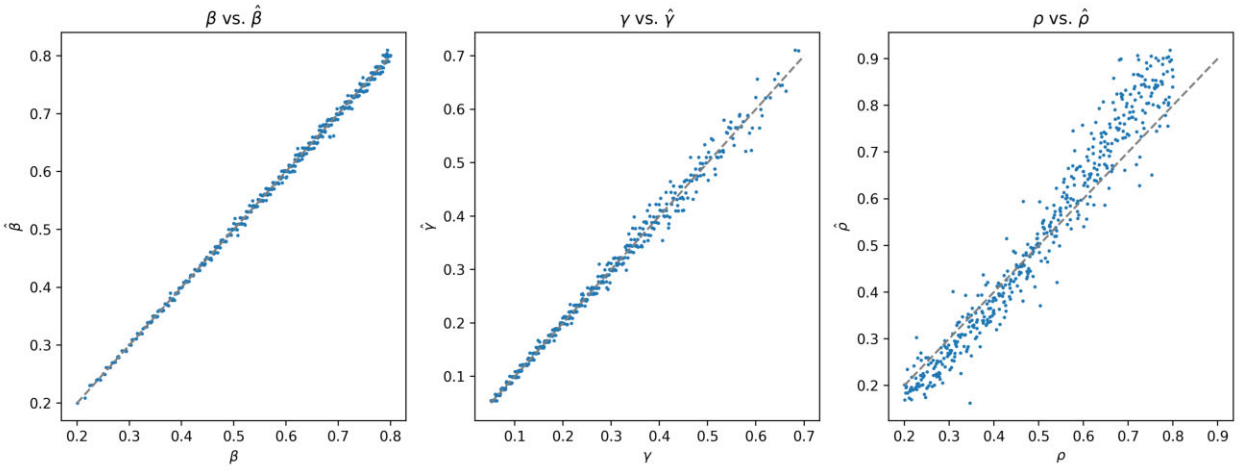

Figure 2. Each plot compares the actual and predicted value of a parameter for many different combinations of the two others. As expected, the estimation of ρ degrades as the actual value gets large. Ultimately, if no one feels sick, behavior does not change and the method cannot pick up ρ

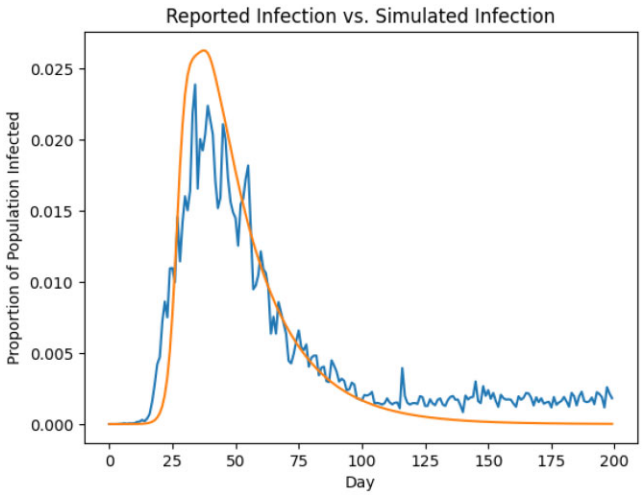

Figure 3. The blue ragged line is the scaled version of the reported daily case count across the New York City metro area, summed across the 44 counties considered here. The orange smooth line is the simulated daily symptomatic case count given our estimated parameters. The scaling factor is chosen to display both lines in a similar magnitude

data about symptomatic infections. The theoretical estimation is derived through (10) and (12). For simulation, we have applied this approach to actual COVID-19 data and derived an estimate of the asymptomatic rate that matches well with the latest estimates obtained via extensive random testing.

While our results are based on the different interaction patterns among asymptomatic and symptomatic populations, it is also possible to distinguish between these two types of spread in two more ways:

1) Distinct infection rates: we can impose dissimilar transmissibility rates in the two types of spread.
2) Differentiated seeding: we can account for several types of spread starting in various locations, perhaps based on local policy or other environmental factors.

Taken together, these factors are a broad set of levers that can influence the emergent behavior of our model, thus potentially enhancing the accuracy of our traSIR model.

Our estimates for β and γ are sharper than for ρ. This is no surprise. Both the transmission rate and the recovery rate have direct influence on the local shape of the infection time series: the first one has a large effect on the ascent phase of the contagion while the other one's impact can be felt most acutely in the descent phase. The impact of the asymptomatic rate ρ is more global and subtle. It can be felt in the speed of the traveling waves and generally operates on longer time scales. TraSIR is able to leverage such information. Credit for our success must also go to sheer luck: An asymptomatic rate of $\sim$30% is almost ideally sized for estimation. As we observed earlier, a rate close to 100% would make the task hopelessly difficult. This leaves open the possibility that other nonlinearities in the system can be exploited to boost accuracy when needed. While fast-changing health policy measures and medical breakthroughs (e.g. vaccination) can present traSIR with major challenges, they also create new windows of opportunity for novel estimation mechanisms. We hope that this work will plant the seeds for exciting new research on the messy, difficult, but fascinating subject of uncovering hidden epidemiological parameters.

Acknowledgements

The authors are pleased to acknowledge that the work reported on in this article was substantially performed using the Princeton Research Computing resources at Princeton University which is a consortium of groups led by the Princeton Institute for Computational Science and Engineering (PICSciE) and Office of Information Technology's Research Computing. M.S. thanks the National Institute of Health (NIH) [R01-GM076275] and the Princeton Catalysis Institute.

Funding

This work was supported in part by National Science Foundation grant Computing and Communication Foundations (CCF) 2006125.

Conflict of Interest

None declared.

References

Ajelli,M. *et al.* (2010) Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models. *BMC Infect. Dis.*, **10**, 190.

Ameri,K. and Cooper,K.D. (2019) A network-based compartmental model for the spread of whooping cough in Nebraska. *AMIA Jt. Summits Transl. Sci. Proc.*, **2019**, 388–397.

American Community Survey (ACS). (2015a) *2011-2015 5-Year ACS Commuting Flows*. Data set. https://www.census.gov/data/tables/2015/demo/metro-micro/commuting-flows-2015.html (9 October 2022, date last accessed).

American Community Survey (ACS). (2015b) *2011-2015 ACS 5-Year Estimates*. Data set. https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2015/5-year.html (9 October 2022, date last accessed).

Anderson,R.M. and May,R.M. (1991) *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

Barabási,A.-L. (2013) Network science. *Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci.*, **371**. http://networksciencebook.com/.

Barrat,A. *et al.* (2008) *Dynamical Processes on Complex Networks*. Cambridge University Press.

Brauer,F. *et al.* (2008) *Mathematical Epidemiology*. *Lectures Notes in Mathematics* 1945. Springer.

Byambasuren,O. *et al.* (2020) Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. *J. Assoc. Med. Microbiol. Infect. Dis. Can., 5, 223–234*. https://doi.org/10.3138/jammi-2020-0030

Connell,C. (2022) *fips2county*. Data set. https://github.com/ChuckConnell/articles/ blob/master/fips2county.tsv (9 October 2022, date last accessed).

Dabke,D.V. and Arroyo,E.E. (2016) Rumors with personality: a differential and agent-based model of information spread through networks. *SIURO*, **9**, 453–467. https://doi.org/10.1137/16s015103

Ding,X. *et al.* (2021) Incorporating dynamic flight network in SEIR to model mobility between populations. *Appl. Netw. Sci.*, **6**, 42.

Dottori,M. and Fabricius,G. (2015) SIR model on a dynamical network and the endemic state of an infectious disease. *Phys. A: Stat. Mech. Appl.*, **434**, 25–35.

Eletreby,R. *et al.* (2020) The effects of evolutionary adaptations on spreading processes in complex networks. *Proc. Natl. Acad. Sci. USA*, **117**, 5664–5670.

Johansson,M.A. *et al.* (2021) SARS-CoV-2 transmission from people without COVID-19 symptoms. *JAMA Netw. Open.*, **4**, e2035057. doi:10.1001/jamanetworkopen.2020.35057

Kermack,W.O. and McKendrick,A.G. (1927) A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A*, **115**, 700–721.

Kerr,C.C. *et al.* (2021) Covasim: an agent-based model of COVID-19 dynamics and interventions. *PLoS Comput. Biol.*, **17**, e1009149. https://doi.org/10.1371/journal.pcbi.1009149

Layton,A.T. and Sadria,M. (2022) Understanding the dynamics of SARS-CoV-2 variants of concern in Ontario, Canada: a modeling study. *Sci. Rep.*, **12**, 2114. https://doi.org/10.1038/s41598-022-06159-x

Li,C. *et al.* (2021) Estimating the prevalence of asymptomatic COVID-19 cases and their contribution in transmission—using Henan province, China, as an example. *Front. Med.*, **8**, 591372.

Li,R. *et al.* (2020) Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, **368**, 489–493. doi: 10.1126/science.abb3221.

Liu,Q.-H. *et al.* (2018) Measurability of the epidemic reproduction number in data-driven contact networks. *Proc. Natl. Acad. Sci. USA*, **115**, 12680–12685.

Ma,Q. *et al.* (2021) Global percentage of asymptomatic SARS-CoV-2 infections among the tested population and individuals with confirmed COVID-19 diagnosis. *JAMA Netw. Open.*, **4**, e2137257.

Martinelli,D. *et al.* (2022) Estimating the proportion of asymptomatic COVID-19 cases in an Italian region with intermediate incidence during the first pandemic wave: an observational retrospective study. *Biomed Res. Int.*, **2022**, 3401566. https://doi.org/10.1155/2022/3401566

Nishiura,H. *et al.* (2020) Estimation of the asymptomatic ratio of novel coronavirus infections (covid-19). *Int. J. Infect. Dis.*, **94**, 154–155.

Nogrady,B. (2020) What the data say about asymptomatic COVID infections. *Nature*, **587**, 534–535.

Poletti,P. *et al.* (2021) Probability of symptoms and critical disease after SARS-CoV-2 infection. *JAMA Netw. Open*, **4**, e211085. arXiv:2006.08471.

Sah,P. *et al.* (2021) Asymptomatic SARS-COV-2 infection: a systematic review and meta-analysis. *Proc. Natl. Acad. Sci. USA*, **118**, e2109229118. https://doi.org/10.1073/pnas.2109229118

Shang,W. *et al.* (2022) Percentage of asymptomatic infections among SARS-CoV-2 omicron variant-positive individuals: a systematic review and meta-analysis. *Vaccines*, **10**, 1049.

Subramanian,R. and Pascual,M. (2021) Quantifying asymptomatic infection and transmission of COVID-19 in New York city using observed cases, serology, and testing capacity. *Proc. Natl. Acad. Sci. USA.*, **118**, 9.

Sun,K. *et al.* (2021) Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science*, 371, eabe2424.

The COVID Tracking Project at The Atlantic. (2022) *Daily data on the COVID-19 Pandemic for the US and Individual States*. Data set. https://api.covidtracking.com/v1/states/daily.csv (9 October 2022, date last accessed).

The New York Times. (2022) *Coronavirus (Covid-19) Data in the United States*. Data set. https://github.com/nytimes/covid-19-data (9 October 2022, retrieved).

Zhan,X.-X. *et al.* (2018) Coupling dynamics of epidemic spreading and information diffusion on complex networks. *Appl. Math. Comput.*, **332**, 437–448.