Story of Two Populations in Epidemics: Is Every Infection Counted?

Van Sy Mai[1] and Richard J. La[2]

[1] ANTD, NIST, Gaithersburg, MD 20899, USA,
vansy.mai@nist.gov
[2] ACMD, NIST, Gaithersburg, MD 20899, USA,
richard.la@nist.gov

Abstract. Many previous studies on epidemic processes assume that most infected nodes that contribute to the spread of infection can be identified and accounted for. But, in some cases, this assumption may be invalid and many infection cases may go undetected. For example, the operators of subsystems in complex systems may be unaware that their systems are compromised by malware or a virus, which may go undetected for a long period. Similarly, an infectious disease can cause widely varying symptoms and some infected individuals may exhibit little to no symptoms. In these scenarios, it would be difficult to identify all infected systems or individuals, which can play a critical role in spreading the infection by malware, virus or disease. For this reason, it is of interest to devise a means of quickly determining the presence of such undetected infection cases in a network. We propose a simple optimization-based approach that can be used to determine whether or not a significant fraction of infection cases are undetected and thus are missing in reported statistics. We present numerical results obtained from a case study using publicly available COVID-19 data from four countries.

1 Introduction

The spread of malware and viruses in computer networks and information systems, infectious diseases, rumors or new product information is often studied using epidemic processes [6, 9, 10]. Many, if not most, of these studies assume that (most) infected nodes can be identified so that the parameters of the epidemic processes, such as reproduction numbers, can be estimated from available data. In some cases, however, such information may not be available or cannot be obtained easily. For example, when subsystems in a network with inadequate security measures are infected by malware or a virus, network managers may be unaware of the infection [7]. Similarly, an infectious disease can cause widely varying symptoms in infected individuals and some individuals may exhibit no obvious symptoms that can be easily identified [13].

Preventing computer malware/viruses or an infectious disease from causing a widespread outbreak requires accurate information about its transmission dynamics at its onset. This task becomes more challenging when it is difficult to

get a clear picture of the underlying epidemic process in a network or society; missing a significant fraction of infection cases in reported numbers at an early stage of an outbreak would likely lead to a costly underestimation of the severity of the problem at a crucial point. This will likely delay urgent measures and policies that should be put in place to curb the spread.

We propose a new approach to determining whether reported numbers of infections are accurate or there are many unreported cases. Our approach borrows insights from (statistical) model selection [11] and is based on a simple observation that infected devices or individuals that go undetected spread the infection differently than those that are detected shortly after infection and likely benefit from remediation (Section 2). A key challenge is that the proposed method must make a determination based solely on the available noisy information.

For our case study, we apply our approach to publicly available COVID-19 data (Section 5). We approximate the time-varying transmissibility of infected individuals with the help of parameterized gamma distributions. We compare the errors between (i) the "best" model with two populations (*detectable* and *undetectable* nodes) and (ii) the "best" model only with a single population (*detectable* nodes). We declare that there is a non-negligible fraction of undetectable infections if (a) the error achieved by the former is considerably less than that of the latter and (b) the estimate of the fraction of unreported infection cases is non-negligible. Numerical results we obtained using the real data for four countries that experienced a COVID-19 outbreak are presented in Section 5; our proposed approach suggests that all four countries likely had a significant fraction of unreported infection cases that contributed to the spread of the disease but are missing in published data.

A few words on notation: the natural number system, the nonnegative real number system, and the positive real number system are denoted by $\mathbb{N} := \{1, 2, \ldots\}$, $\mathbb{R}_+ := [0, \infty)$, and $\mathbb{R}_{++} := (0, \infty)$, respectively. We will use boldface letters or symbols to denote vectors.

2 Setup

As mentioned earlier, our goal is to quickly determine, on the basis of a time series consisting of the numbers of confirmed reported infection cases over a period, whether the reported numbers accurately reflect the true numbers or they are missing many undetected, hence unreported infection cases. In order to make progress, we adopt the well-known susceptible-infected-recovered (SIR) model to capture the transmission dynamics.

The system or population consists of P (nondescript) subsystems or individuals, which we refer to as *nodes*. We assume that P is large and denote the set of P nodes by $\mathcal{V}$. The set $\mathcal{V}$ can be partitioned into two populations – *detectable* or *undetectable*: when a detectable node (DN) is infected, it exhibits symptoms associated with the infection and is identified as 'infected'. By contrast, when an undetectable node (UN) becomes infected, it goes undetected and thus is not identified as 'infected'. For example, when a computer or device is infected with a virus, it may run a malicious process in the background or conduct a network

probe, looking for vulnerabilities. Such activities can be detected by appropriate network monitoring and analysis tools in a suitably protected network. On the other hand, a poorly monitored network, which represents an UN, may not have the capability to detect an infection, which will go uninvestigated as a result. In the case of an infectious disease, a DN is an individual who, when infected, will exhibit identifiable symptoms and be reported as a confirmed case. In contrast, an infected UN is either asymptomatic or does not manifest her infection with serious symptoms, and hence is not reported as a confirmed case.

Let $p_U \in [0, 1)$ be the fraction of undetectable nodes in $\mathcal{V}$, which is unknown beforehand, and define $P_U := P \cdot p_U$ and $P_D := P - P_U = P(1 - p_U)$. We aim to devise a method for determining whether $p_U \approx 0$ or not. Our approach, which is described in detail in subsequent sections, is based on the following observation: when DNs are infected and exhibit symptoms, they will be identified and isolated in order to stop the spread of infection. As a result, most of the transmissions will take place before its infection is detected. For instance, according to the U.S. Centers for Disease Control and Prevention (CDC), approximately 50 percent of COVID-19 transmissions occur prior to the onset of symptoms (presymptomatic) [1]. On the other hand, when UNs are infected, because they are unaware of the infection, they continue to spread the infection to others over longer periods. We take advantage of this discrepancy in the manner in which DNs and UNs spread infection to devise our scheme.

3 Approach and Model

The key aspect of the discrepancy in the spread of infection by DNs and UNs, which we hope to exploit in our approach, can come in different forms. For example, when we consider the spread of malware/viruses in a network, even when they are infected, the devices equipped with the latest anti-malware/virus software will likely detect the infection sooner than devices with inadequate protection. For this reason, their infection periods will vary widely. In the case of an infectious disease, infected individuals with more pronounced symptoms will likely be isolated or hospitalized, preventing them from further spreading the diseases, whereas infected yet asymptomatic individuals will continue to spread the disease, possibly with lower infectiousness.

This observation tells us that in the event that a DN becomes infected and its infection is detected, it will be removed from the network so that its infection does not spread and is put back in the network only after it is no longer infectious upon recovery/repair. Therefore, once detected, even though the infected DN may remain infectious, its ability to transmit the infection will be limited. For this reason, we find it convenient to differentiate the rate at which infected nodes can spread its infection from its infectiousness and call it the *transmissibility*.

3.1 Transmissibility Function

Following an approach similar to [8], we approximate the time-varying transmissibility of an infected node with the help of some function: suppose that

a DN (resp. UN) is exposed to the etiological agent at time $t_0 \in \mathbb{R}_+$. Then, the transmissibility of the infected node at time $t \geq t_0$ is given by $f_D(t-t_0)$ (resp. $f_U(t-t_0)$). Moreover, the *total* transmissibility over time, which is given by $\int_{\mathbf{R}_+} f_D(\tau)\, d\tau$ or $\int_{\mathbf{R}_+} f_U(\tau)\, d\tau$, is related to the expected number of transmissions or the (basic) reproduction number of an infected DN or UN, respectively [8].

3.2 Discrete-Time Approximation of Transmissibility Function

For our study, we adopt a discrete-time model and approximate the evolution of the total number of infection cases using the following equations, which can be viewed as a discrete-time approximation of the von Foerster equation used in [8] with two populations – DNs and UNs: for each $n \in \mathbb{N}$, let $N_D(n)$ (resp. $N_U(n)$) be the number of DNs (resp. UNs) newly infected at time $n \in \mathbb{N}$ and $\mathbf{N}(n) := (N_D(n), N_U(n))$. Similarly, let $I_D(n) := \sum_{k=1}^{n} N_D(k)$ (resp. $I_U(n) := \sum_{k=1}^{n} N_U(k)$) be the total number of infected DNs (resp. UNs) up to and including time n, and $\mathbf{I}(n) := (I_D(n), I_U(n))$. For each $n \in \mathbb{N}$, define $I_T(n) := I_D(n) + I_U(n)$, $q(n) := I_T(n)/P$, $q_D(n) := I_D(n)/P_D$, and $q_U(n) := I_U(n)/P_U$.

Suppose that there is finite $d_{\max} \in \mathbb{N}$ such that the discrete-time transmissibility functions $\mathcal{K}_D(k)$ and $\mathcal{K}_U(k)$ for an infected DN and UN, respectively, are negligible for $k > d_{\max}$. Under this assumption, we can view the discrete-time transmissibility functions as finite sequences $\mathcal{K}_D = \big(K_D(n) : n = 1, 2, \ldots, d_{\max}\big)$ and $\mathcal{K}_U = \big(K_U(n) : n = 1, 2, \ldots, d_{\max}\big)$. For notational convenience, for each $n \in \mathbb{N}$, we denote the finite sequences $\big(N_D(n') : n' = n - d_{\max} + 1, \ldots, n\big)$ and $\big(N_U(n') : n' = n - d_{\max} + 1, \ldots, n\big)$ by $\mathcal{N}_D(n)$ and $\mathcal{N}_U(n)$, respectively.

The aggregate transmissibility of the previously infected nodes at time n is equal to

$$N_T(n) = \mathcal{N}_D(n-1) \circledast \mathcal{K}_D + \mathcal{N}_U(n-1) \circledast \mathcal{K}_U, \tag{1}$$

where $\mathcal{N}_D(n-1) \circledast \mathcal{K}_D := \sum_{k=1}^{d_{\max}} N_D(n-k)\ K_D(k)$ and $\mathcal{N}_U(n-1) \circledast \mathcal{K}_U := \sum_{k=1}^{d_{\max}} N_U(n-k)\ K_U(k)$. Using the aggregate transmissibility, we can compute the expected numbers of newly infected DNs and UNs at time n given by

$$N_D(n) = N_T(n)\ (1 - p_U)\ (1 - q_D(n-1)) \quad \text{and} \tag{2a}$$

$$N_U(n) = N_T(n)\ p_U\ (1 - q_U(n-1)). \tag{2b}$$

Recall that $q_D(n-1) = I_D(n-1)/P_D$ and $q_U(n-1) = I_U(n-1)/P_U$. Substituting these in (2a) and (2b) and taking their ratio tells us $N_U(n)/N_D(n) = (P_U - I_U(n-1))/(P_D - I_D(n-1)) = (P_U - I_U(n))/(P_D - I_D(n)) = N_U(n+1)/N_D(n+1)$ for all $n \in \mathbb{N}$ and, thus, $I_U(n)/I_D(n)$ is constant. In view of this observation, it is reasonable to expect the ratio $I_U(n)/I_D(n)$ to remain close to $p_U/(1-p_U)$, and we make a simplifying assumption $q_D(n) = q_U(n) = q(n)$ for all $n \in \mathbb{N}$.

This assumption allows us to simplify equations (1) and (2a) as follows.

$$N_T(n) = \mathcal{N}_D(n-1) \circledast \Big(\mathcal{K}_D + \tfrac{p_U}{1-p_U}\mathcal{K}_U\Big) \quad \text{and}$$

$$N_D(n) = \mathcal{N}_D(n-1) \circledast \big(((1-p_U)\mathcal{K}_D + p_U\mathcal{K}_U)\big(1 - q(n-1)\big)\big) \tag{3}$$

Thus, if we knew p_U and the transmissibility functions $\mathcal{K}_D$ and $\mathcal{K}_U$, we would be able to predict the evolution of $N_D(n)$ and $N_U(n)$, $n \in \mathbb{N}$. This suggests that we can approach our problem from the perspective of *model selection*, where we wish to identify the functions $\mathcal{K}_D$ and $\mathcal{K}_U$ as well as the parameter p_U.

Unfortunately, without imposing additional assumptions, accurate estimation of the parameter p_U and the functions $\mathcal{K}_D$ and $\mathcal{K}_U$ requires a large number of samples, which may not be available, especially at the onset of an epidemic when we wish to make a determination. Oftentimes, this difficulty can be mitigated by first choosing a set of candidate functions and finding the functions that best match the available observations/samples. This is the approach we take, which is explained in detail in subsequent sections.

4 Parameter Estimation and Model Selection

Suppose that $\mathbb{K} \subset \mathbb{R}_+^{d_{\max}}$, i.e., a set of nonnegative sequences of length $d_{\max}$, to which we assume $\mathcal{K}_D$ and $\mathcal{K}_U$ belong. Our goal is to identify the "best" sequences $\mathcal{K}_D^*$ and $\mathcal{K}_U^*$ in $\mathbb{K}$ and parameter p_U^*, which minimize the discrepancy between the available observations and those predicted by the model, which is measured by some performance measure. In theory, this problem can be formulated as an optimization problem. But, as mentioned earlier, because this optimization problem has $1+2d_{\max}$ optimization variables, unless $d_{\max}$ is small, which may not be the case in general, it would require many samples to find good estimates.

To address this issue, we propose to consider a family of sequences that are parameterized by a few parameters denoted by $\boldsymbol{\xi}$. We assume that $\boldsymbol{\xi}$ belongs to a parameter set $\boldsymbol{\Xi}$. For fixed parameter set $\boldsymbol{\Xi}$, define $\mathbb{K}^{\boldsymbol{\Xi}}$ ($\subsetneq \mathbb{K}$) to be set of parameterized sequences. Since we approximate the time-varying transmissibility of infected DNs and UNs using parameterized sequences in $\mathbb{K}^{\boldsymbol{\Xi}}$, we only need to estimate the parameters of the best sequences as well as the parameter p_U, which requires fewer samples. To this end, we employ an optimization problem described later in this section.

Clearly, the quality of the solution we find by solving an appropriate optimization problem will depend on the allowed set $\mathbb{K}^{\boldsymbol{\Xi}}$. Thus, some domain knowledge may be necessary to identify a suitable set of parameterized sequences that can be used to approximate the time-varying transmissibility. In Section 5, we adopt a family of scaled gamma kernels for the case study of COVID-19.

• **Predicted time series for fixed parameters:** Suppose that $\mathcal{T} = \big(T(n) : n = 1, 2, \ldots, n_{\max}\big)$ is the available finite time series of length $n_{\max} \geq 1$, where $T(n)$ is the total numbers of reported infections up to and including time n. Given a feasible solution $\mathbf{x} := (\tilde{p}_U, \tilde{\boldsymbol{\xi}}_D, \tilde{\boldsymbol{\xi}}_U) \in [0,1) \times \boldsymbol{\Xi}^2 =: \mathbb{X}$, where $\tilde{\boldsymbol{\xi}}_D$ and $\tilde{\boldsymbol{\xi}}_U$ are the parameters of $\mathcal{K}_D$ and $\mathcal{K}_U$, respectively, we construct the following two sequences, $\tilde{\mathcal{N}}_D^{\mathbf{x}}$ and $\tilde{\mathcal{I}}_D^{\mathbf{x}}$: let $n_1 = \min\{n \in \mathbb{N} \mid T(n) \geq \tau_{\min}\}$, where $\tau_{\min} > 0$ is a minimum infection number parameter we select. Without loss of generality, we assume $n_1 = 1$. We then set $q(n_1) = T(n_1)/(P \times \tilde{p}_U)$, $\tilde{N}_D^{\mathbf{x}}(n_1) = T(n_1)$ and $\tilde{N}_D^{\mathbf{x}}(n') = 0$ for all $n' = n_1 - d_{\max}+1, \ldots, n_1 - 1$. Finally, starting with $n = n_1 + 1$, we use equation (3) to recursively generate the *predicted* numbers of new cases

$\tilde{N}_D^{\mathbf{x}}(n)$ for $n \in \{n_1+1, n_1+2, \ldots, n_{\max}\}$, with the initial condition $\tilde{\mathcal{N}}_D^{\mathbf{x}}(n_1) = (\tilde{N}_D^{\mathbf{x}}(n') : n' = n_1 - d_{\max}+1, \ldots, n_1)$. The predicted time series of the total number of reported infections is then given by $\tilde{\mathcal{I}}_D^{\mathbf{x}} = \big(\tilde{I}_D^{\mathbf{x}}(n) = \sum_{k=n_1}^{n} \tilde{N}_D^{\mathbf{x}}(k) : n = n_1, n_1+1, \ldots, n_{\max}\big)$.

• **Objective function:** To find the "best" parameters $\mathbf{x}^\star$, we use the mean squared error (MSE) as a metric to measure the distance between the given time series $\mathcal{T}$ and a predicted times series $\tilde{\mathcal{I}}_D^{\mathbf{x}}$:

$$\mathcal{D}\big(\mathcal{T}, \tilde{\mathcal{I}}_D^{\mathbf{x}}\big) = \sum_{n=n_1}^{n_{\max}} \big(T(n) - \tilde{I}_D^{\mathbf{x}}(n)\big)^2 \tag{4}$$

1) **Two-population model (TPM):** In order to find optimal parameters $\mathbf{x}^\star$ that best approximate the given time series $\mathcal{T}$ with respect to the MSE in (4) with two populations – DNs and UNs – we solve the following constrained optimization problem:

Opt. prob. for parameter estimation with two populations

$$\text{minimize}_{\mathbf{x}\in\mathbb{X}} \quad \mathcal{D}\big(\mathcal{T}, \tilde{\mathcal{I}}_D^{\mathbf{x}}\big) \tag{5}$$

We denote by $\mathbb{X}^\star$ the optimal set, i.e., the set of optimal points of (5), and $v^\star$ is the optimal value, i.e., $v^\star = \mathcal{D}\big(\mathcal{T}, \tilde{\mathcal{I}}_D^{\mathbf{x}^\star}\big)$, where $\mathbf{x}^\star \in \mathbb{X}^\star$ is an optimal point.

2) **Single-population model (SPM):** To answer our question of whether or not there is a significant number of undetected infections that go unreported, we also solve the following related optimization problem only with a single population of DNs. First, from (3), when there are no undetected cases, we approximate the epidemic dynamics only with DNs using

$$N_D(n) = \mathcal{N}_D(n-1) \circledast \mathcal{K}_D\big(1 - q(n-1)\big), \quad n \in \mathbb{N}, \tag{6}$$

where $\mathcal{K}_D \in \mathbb{K}^{\boldsymbol{\Xi}}$, and obviously the parameters $\tilde{p}_U$ and $\tilde{\boldsymbol{\xi}}_U$ play no role in the epidemic dynamics, which is clear from (6). Define $\mathbb{X}_S := \{\mathbf{x}' \in \mathbb{X} \mid p'_U = 0\} \subsetneq \mathbb{X}$.

Opt. prob. for parameter estimation with a single population

$$\text{minimize}_{\mathbf{x}\in\mathbb{X}_S} \ \mathcal{D}\big(\mathcal{T}, \tilde{\mathcal{I}}_D^{\mathbf{x}}\big) \tag{7}$$

We denote the optimal set and the optimal value of (7) by $\mathbb{X}_S^\star$ and $v_S^\star$, respectively. Because $\mathbb{X}_S \subsetneq \mathbb{X}$, we have $v^\star \leq v_S^\star$.

4.1 Model Selection

Determining whether or not there is a considerable fraction of UN infections based on the optimal values of (5) and (7), can be cast as a model selection problem. Loosely speaking, if all optimal points $\mathbf{x}^\star$ in $\mathbb{X}^\star$ have $p_U^\star \approx 0$, then it likely suggests that at most a small fraction of infections may be undetected

or unreported. On the other hand, if $p_U^\star$ is not close to zero for some optimal points in $\mathbb{X}^\star$, it is still unclear if we should conclude that there are a significant number of unreported cases. This is especially the case if $v^\star \approx v_S^\star$ and the TPM does not significantly improve the approximation error (measured by the MSE) compared to the SPM. This is because the minor improvement in MSE may be due to 'overfitting', which is a well-known issue in model selection [11].

To cope with this problem, we borrow an idea from model selection theory, specifically information criteria [11]: in order to avoid overfitting, an information criterion, e.g., Akaike information criterion or Bayesian information criterion, is often used in model selection to carry out a trade-off between the accuracy of a chosen model and the number of model parameters. Unfortunately, these information criteria are not appropriate for our problem.

Instead, following a similar intuition behind information criteria, we base our decision on the relative improvement in the MSE we get using the TPM over the SPM: define $\theta^+ := (v_S^\star - v^\star)/v_S^\star$ to be the relative improvement in MSE. We conclude that a significant fraction of cases come from UNs if (a) $\theta^+ \geq \theta_{th}$ and (b) $p_U^\star \geq p_{th}$, where θ_{th}, $p_{th} \in (0, 1)$ are thresholds we select.

5 Case Study Using Real Data: COVID-19

We carried out a case study using publicly available data for COVID-19. One of surprising aspects of COVID-19 which made its suppression more difficult, especially at the onset, is that many of infected individuals showed no serious outward symptoms, allowing them to unwittingly spread the infection and making contact tracing less effective. The reported estimates of asymptomatic cases vary widely due to several difficulties, and the current best CDC estimate as of August 2021 is approximately 30 percent.[3] Moreover, according to the CDC estimate, infected asymptomatic individuals (AIs) are about 75 percent as infectious as symptomatic individuals (SIs), posing a serious challenge [1].

We provide some numerical results that we obtained using the data downloaded from a public website that keeps track of the numbers of COVID-19 cases worldwide [2]. In this case study, reported cases come from infected DNs, whereas UNs do not report when infected, due to lack of serious symptoms or testing. Note that, in our setting, UNs may include SIs with mild symptoms which did not require testing/reporting, while the DNs may include some reported AIs confirmed through testing. However, most asymptomatic cases are assumed unreported. Thus, the fraction of UNs estimated using our approach will likely be larger than the fraction of AIs, denoted by p_A, as it includes unreported SIs.

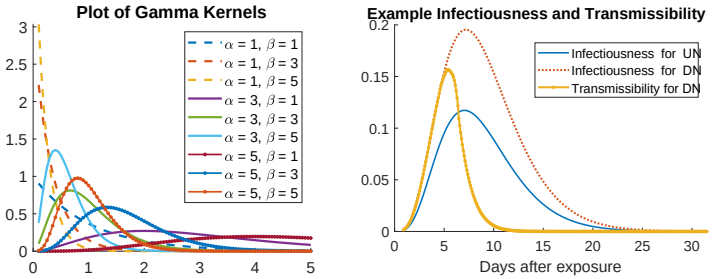
• **Family of Gamma Distributions** – In our study, we approximate the time-varying transmissibility functions of infected DNs and UNs using a family of discrete-time gamma kernels: for given parameters $\boldsymbol{\xi} = (\alpha, \beta, r) \in \mathbb{R}^3_{++} (= \boldsymbol{\Xi})$, the scaled discrete-time gamma kernel $\mathcal{K}^{\boldsymbol{\xi}}$ is given by a sequence $\left(K^{\boldsymbol{\xi}}(n) : n =\right.$

[3] The CDC estimate of asymptomatic cases has changed over time. For instance, this estimate was 40 percent in November 2020.

$1, 2, \ldots, d_{\max}\big)$, where

$$K^{\boldsymbol{\xi}}(n) = r\frac{\beta^{\alpha}}{\Gamma(\alpha)}n^{\alpha-1}e^{-\beta n}, \quad n \in \{1, 2, \ldots, d_{\max}\}, \tag{8}$$

Γ is the gamma function, and r is a scaling constant, which in our problem captures the basic reproduction number. We denote the (discrete-time) transmissibility function of DNs (resp. UNs) of the form in (8) by $\mathcal{K}_D$ (resp. $\mathcal{K}_U$).



(a) (b)

Fig. 1. (a) Family of gamma kernels for different parameter values, and (b) example infectiousness and transmissibility. In Fig. 1(b), we assume that the incubation period is 3-5 days and an infected DN isolates herself starting on day 5 or 6.

As shown in Fig. 1(a), gamma kernels with different values of parameters can be used to approximate functions with varying shapes and supports. Moreover, some of these gamma kernels closely resemble the example infectiousness functions reported in [8, Fig. 1], suggesting that gamma kernels are indeed good candidate functions for approximating the transmissibility functions. Since gamma kernels are probability density functions, their integrals are always equal to one. For this reason, we scale them to reflect the unknown average number of transmissions, i.e., reproduction numbers.

An exposed individual may become infectious as early as a few days into incubation period and remains contagious for several days after showing symptoms, followed by a slow decline in infectiousness [3]. However, once an infected DN shows symptoms and isolates herself, she may not transmit the disease much, indicated by low transmissibility. For an infected UN, on the other hand, her transmissibility will remain high, albeit slightly lower peak infectiousness than that of a DN as suggested in [1], for a longer period throughout her infectious period. These observations suggest that we should look for a gamma kernel with a wider support for the transimissiblity function of UNs compared to that of DNs. Examples of infectiousness and transmissibility are shown in Fig. 1(b).

5.1 COVID-19 Outbreak Data Used in the Study

For our study, among many countries that experienced an outbreak, we selected four countries – France, Spain, the United Kingdom (U.K.), and the United

States (U.S.) – and applied our approach using the publicly available daily numbers of reported cases in each country at the beginning of the pandemic [2].

Since the infectious period of an infected individual with moderate symptoms is believed to last about 7 to 12 days [4] and we wanted to consider the time series before widespread implementation of public health measures and policies in response, e.g., social distancing and face mask mandates, we considered the time series for a period of 30 days in early stages before the number of reported confirmed cases reached 10,000 cases in the countries.[4] The reported numbers we considered are for the period of 30 days from February 15, 2020 to March 15, 2020, except for the U.K., for which we considered the period of February 20, 2020 through March 20, 2020. The reason for this shift in the studied period for the U.K. is that there were 1,263 confirmed cases in the U.K. on March 15, 2020, which is far lower than the other three countries and our approach produced 500 estimates of p_U which were more spread out, indicating that there may not have been enough reported cases to produce consistent solutions.

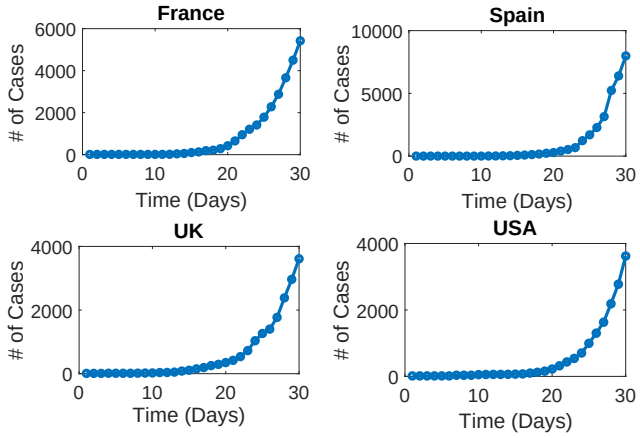


Fig. 2. Number of reported COVID-19 cases in France, Spain, the U.K., and the U.S.

The total number of reported cases over the periods under consideration for each country is shown in Fig. 2. It is clear that, despite some variations, the number of confirmed cases increases in a similar fashion. This is unsurprising because the number of infections is expected to grow exponentially at the onset of an epidemic before public health measures are introduced to curb the spread.

5.2 Summary of Numerical Results

Before presenting the numerical results, we briefly describe a technical difficulty that arises due to the fact that the constrained optimization problems in (5) and

[4] Once widespread public health measures are in place, the spread dynamics are likely to change significantly with different parameters. For this reason, we considered the infection numbers in early stages.

(7) are nonconvex. In fact, our numerical studies suggest that there are many local optima, many of which are of poor quality. For this reason, we randomize the initial solution for parameter estimation, solve the optimization problems in equations (5) and (7) 500 times, and choose the best solution(s) out of the 500 solutions we produce as our best estimates of the optimal points.

We solved the optimization problems using the `fmincon` function built in MATLAB as well as `simulannealbnd` (simulated annealing) and `particleswarm` (particle swarm optimization).[5] Both `simulannealbnd` and `particleswarm` took much longer and produced solutions of comparable quality. Here we only report the solutions obtained using the `fmincon` function.

	France	Spain	UK	USA
$v^\star$	4.60×10^4	2.14×10^5	1.80×10^4	7.83×10^3
$v_S^\star$	1.15×10^5	3.88×10^5	4.78×10^4	1.86×10^4
$p_U^\star$	0.604	0.662	0.480	0.779
θ^+	0.60	0.45	0.62	0.58

Table 1. Summary of Numerical Results.

Table 1 summarizes the numerical results, including the minimum MSEs achieved by the best solution out of the 500 solutions for the two models (SPM and TPM) as well as the corresponding θ^+. It also provides the best estimate $p_U^\star$ of the fraction of UNs. It is clear from the reported numbers, especially θ^+, that the TPM achieves a much smaller MSE than the SPM. Together with non-negligible estimates $p_U^\star$, this suggest that there may be many more cases than reported, likely including a significant fraction of asymptomatic cases that went unreported in all four countries. This is consistent with several recent studies (e.g., [12]). In particular, the current CDC estimate [5] suggests that only 1 in 4.2 or roughly 23.8 percent of COVID-19 infections might have been reported.[6] This is close to our estimate of 22.1 percent for the U.S.

Two-pop. model	p_U	0.6039	0.6114	0.4861	0.0899	0.2229
	MSE	4.60×10^4	4.81×10^4	5.85×10^4	7.29×10^4	9.21×10^4
Single-pop. model	MSE	1.15×10^5	1.15×10^5	1.15×10^5	1.15×10^5	1.15×10^5

Table 2. Five best estimates of p_U and MSEs for SPM and TPM (France).

Fig. 3 shows the distribution of the 5 best estimates of p_U (top 1 percent) for each country from the 500 solutions we obtained. To obtain these values, we first ordered the 500 solutions we generated using the TPM, by increasing value of MSE, and took the first 5 solutions with the 5 smallest MSEs.

As we can see from the figure, the 5 best estimates of p_U are concentrated around the best estimate for three countries – Spain, the U.K., and the U.S.

[5] Any mention of commercial products in this paper is for information only and is not intended for an endorsement or recommendation by NIST.

[6] This estimate is for the period of February 2020 through May 2021.

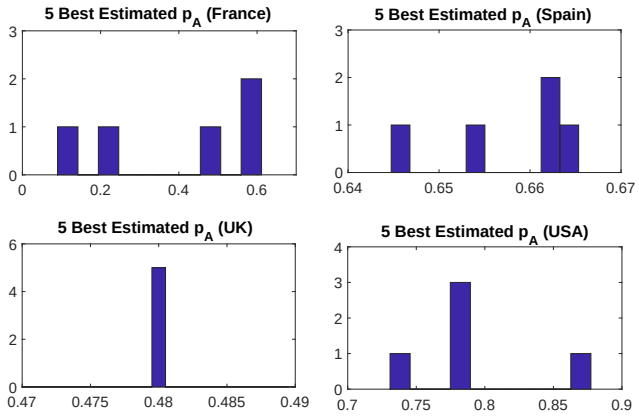


Fig. 3. Histogram of the five best estimated p_A (out of 500).

In fact, the best 10 estimates for the U.K. are identical with the same MSE, suggesting that the obtained estimate $p_U^\star = 0.48$ is likely a good estimate.

In contrast, for France only one other estimate is close to the best estimate, and two estimates are in fact considerably smaller. This is due to the difficulty explained at the beginning of this subsection that, in some cases, the optimization in (5) has many local optima and finding good solutions close to optimal point(s) is difficult. As a result, only a few solutions out of the 500 solutions we generated are good-quality solutions near optimal point(s); starting with the third best solution, the quality of the solution begins to deteriorate quickly with rapidly increasing MSE. This is illustrated in Table 2. On the other hand, the MSE achieved by the 5 best solutions of (7) for the SPM is identical. This is due to the fact that the parameter estimation for SPM is much easier than that for TPM; there are only 3 unknown parameters (α_S, β_S, r_S) to estimate for SPM compared to 7 unknown parameters in TPM.

As explained in Section 4, because our TPM treats *unreported* SIs as a part of UNs, the estimates $p_U^\star$ tend to overestimate p_A. For this reason, it is not our claim that $p_U^\star$ is an accurate estimate of p_A. Instead, large values of $p_U^\star$ together with significant relative improvements (θ^+) suggest that indeed there might be two different populations – DNs and UNs – that drove the epidemic dynamics at the onset of the pandemic, with the possibility of many asymptomatic cases.

Finally, we would like to mention that solving the two optimization problems in (5) and (7) 500 times for a country, using the MATLAB (version 2019b) `fmincon` optimization function takes less than 15 minutes on a laptop with an Intel Core i7-2630QM CPU @ 2.0GHz and 8 GB RAM. Hence, the proposed approach is computationally efficient.

6 Conclusion

We studied the problem of quickly determining whether or not there are many unreported cases of infection at the onset of an epidemic. Having an accurate

answer to this question in early stages of an epidemic, will likely help the health experts devise better and more timely public health measures and policies. Our approach examines whether or not a TPM with both DNs and UNs can better predict the observed numbers than a SPM with only DNs. Our numerical studies using publicly available COVID-19 data in four countries suggest that a significant fraction of total infections in these countries may not have been reported, and the true numbers may be considerably larger.

References

1. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html.
2. https://www.worldometers.info/coronavirus/.
3. https://www.health.harvard.edu/diseases-and-conditions/if-youve-been-exposed-to-the-coronavirus.
4. https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html.
5. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html.
6. Noam Berger, Christian Borgs, Jennifer T. Chayes, and Amin Saberi. On the spread of viruses on the internet. In *the proc. of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, January 2005.
7. Leyla Bilge and Tudor Dumitras. Before we knew it: an empirical study of zero-day attacks in the real world. In *the proc. of ACM Conference on Computer and Communications Security*, pages 833–844, October 2012.
8. Christophe Fraser, Steven Riley, Roy M. Anderson, and Neil M. Ferguson. Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 101(16):6146 – 6151, April 2004.
9. José Luis Iribarren and Esteban Moro. Branching dynamics of viral information spreading. *Physical Review E*, 84(046116), 2011.
10. Richard J. La. Cascading failures in interdependent systems: impact of degree variability and dependence. *IEEE Transactions on Network Science and Engineering*, 5(2):127–140, 2018.
11. H. Linhart and W. Zucchini. *Model Selection, Wiley Series in Probability and Statistics*. Wiley, 1986.
12. Daniel P. Oran and Eric J. Topol. Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review. *Annals of Internal Medicine, M20-3012*, June 2020.
13. Rahul Subramanian, Qixin He, and Mercedes Pascual. Quantify asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity. *Proceedings of the National Academy of Sciences of the United States of American*, 67(026126), 2021.