Data-driven Identification of Number of Unreported Cases for COVID-19: Bounds and Limitations

Ajitesh Srivastava
Viktor K. Prasanna
ajiteshs@usc.edu
prasanna@usc.edu
University of Southern California
Los Angeles, CA, USA

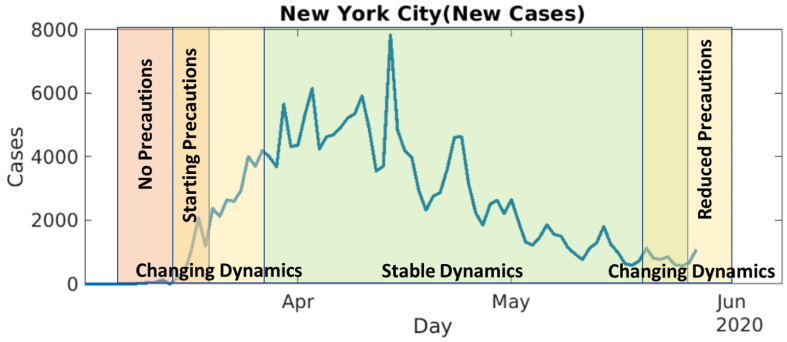


Figure 1: The social distancing phase allows us to model COVID-19 in a way when the effect of latent unreported/asymptomatic cases can be reliably observed.

ABSTRACT

Accurate forecasts for COVID-19 are necessary for better preparedness and resource management. Specifically, deciding the response over months or several months requires accurate long-term forecasts which is particularly challenging as the model errors accumulate with time. A critical factor that can hinder accurate long-term forecasts, is the number of unreported/asymptomatic cases. While there have been early serology tests to estimate this number, more tests need to be conducted for more reliable results. To identify the number of unreported/asymptomatic cases, we take an epidemiology data-driven approach. We show that we can identify lower bounds on this ratio or upper bound on actual cases as a factor of reported cases. To do so, we propose an extension of our prior heterogeneous infection rate model, incorporating unreported/asymptomatic cases. We prove that the number of unreported cases can be reliably estimated only from a certain time period of the epidemic data. In doing so, we construct an algorithm called Fixed Infection Rate method, which identifies a reliable bound on the learned ratio. We also propose two heuristics to learn this ratio and show their effectiveness on simulated data. We use our approaches to identify the upper bounds on the ratio of actual to reported cases for New York City and several US states. Our results demonstrate with high confidence that the actual number of cases cannot be more than 35 times in New York, 40 times in Illinois, 38 times in Massachusetts and 29 times in New Jersey, than the reported cases.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; • **Computing methodologies** → **Model development and analysis**.

KEYWORDS

COVID-19, epidemiological modeling, unreported cases, model learning

ACM Reference Format:
Ajitesh Srivastava and Viktor K. Prasanna. 2018. Data-driven Identification of Number of Unreported Cases for COVID-19: Bounds and Limitations. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

During the current COVID-19 pandemic, researchers have attempted to estimate the number of cases that are not being reported using antibody tests [5]. This number is useful as it dictates the number of susceptible individuals, which in turn affects the long-term dynamics of the epidemic.

We take a data-driven approach to model the existence of unreported cases in terms of probability of a case being reported. Due to a long period of social distancing, the infection dynamics are 'stable',i.e, the parameters that drive the number of cases can be assumed to be constant over the period. This is unlike the earlier phase when the world had just started taking precautions during which a single model with fixed parameters would not have been able to explain the trends. Using the data from this "stable" phase (see Figure 1) of social distancing phase and before the precautions are reduced, we may be able to observe the effect of unreported cases. We demonstrate that the probability of reporting can be reliably obtained only from certain parts of the time-series. This in turn provides an estimated upper bound on the number of total actual cases as a factor of number of reported cases. Particularly, we prove that the probability of reporting has a negligible effect on the trend of reported cases in the initial part of the epidemic. Therefore, during that period, we cannot reliably learn the reporting probability. On the other hand, we also prove that learned probability is not reliable using only the later phase of the epidemic. Thus, there is a certain time interval over which the learned bound on reporting probability is reliable. We leverage the fact that reporting probability has negligible effect on the initial part of the timeseries and significant impact in the later part to construct an algorithm termed Fixed Infection Rate method. Our method can guarantee that the obtained upper bound is close to the true upper-bound. We also propose two heuristics that attempt to learn this upper-bound without any guarantee. While we can also attempt to identify this bound without relying on a 'stable' phase using adaptive models [11], it will introduce more hyperparameters making our estimation less reliable.

We are learning a lower bound on reporting probability (and correspondingly, upper bound on the actual cases) because we can only measure the combined effect of probability of reporting and complete isolation (see Section 3.1). This complete isolation is different from reducing social interactions. Reduced social interactions reduces the probability of a randomly selected infected person affecting a randomly selected susceptible person. On the other hand, complete isolation implies that a part of the population is removed and does not participate in the epidemic, effectively reducing the population by a constant factor. Since this factor is not known, we can only obtain a lower bound on reporting probability or an upper bound on the total cases as a factor of reported cases.

We proceed with an extension of the model proposed in [11] which has been shown to perform accurate forecasts. We have previously used a preliminary version of this model in the DARPA Chikungunya forecasting challenge [3], where we were one of the winners [2]. However, our approach for identifying the right data to reliably learn reporting probability may be applicable to other epidemiological models as well. Our contributions are the following:

- We propose an extension of our prior heterogeneous infection rate model that incorporates unreported/asymptomatic cases in the form of a parameter that measures the ratio of reported cases to actual number of cases.
- We prove that a bound on number of unreported cases can be reliably estimated only from certain data.
- We propose Fixed Infection Rate Learning, an algorithm that leverages the effect of data on the model parameters to reliably identify a lower bound on reporting probability (and correspondingly, upper bound on actual cases as a factor of reported cases).
- We also propose two heuristics – Non-linear Incremental Learning and Non-linear Curve Fitting, that attempt to learn a lower bound on reporting probability, but do not provide reliability guarantees.
- On simulated data, we show that our proposed method and proposed heuristics are accurately able to retrieve the ratio of reported to actual cases.
- We use our approaches to identify the lower bounds on the ratio of reported to actual cases for New York City and several US states.Our results demonstrate with high confidence that the actual number of cases are cannot be more than 35 times in New York, 40 times in Illinois, 38 times in Massachusetts and 29 times in New Jersey, than the reported cases.

2 RELATED WORK

2.1 Modeling Unreported Cases

Several works in the literature [8–10] have attempted to model unreported cases by adding states such as asymptomatic and unreported to the Susceptible-Infected-Removed (SIR) model [6]. Magal and Webb [10] propose a methodology for SIR model, that can determine the probability of reporting. This approach assumes that the 'turning point', i.e., the time at which the number of new cases peaks, is known. Ducrot et. al. [8] propose a method for identification of unreported cases from reported cases when the model parameters satisfy certain properties in an extension of SIR model. Liu et. al. [9] use a similar model but do not discuss the learnability of parameters related to asymptomatic and unreported cases.

2.2 The SI-kJα Model

In [11], we proposed the SI-kJα model for the spread of a virus like COVID-19 across the world which captures (i) temporally varying infection rates (ii) arbitrary regions, and (iii) human mobility patterns. Within every region (hospital/city/state/country), an individual can exist in either one of two states: susceptible and infected. A susceptible individual gets infected when in contact with an infected individual at a rate depending on when that individual got infected, i.e., rate of infection is β_1 for an individual infected between $t-1$ and $t-J$, β_2 for an individual infected between $t-J$ and $t-2J$, and so on, thus resulting in k sub-states of infection. J is a hyperparameters introduced for a smoothing effect to deal with noisy data. It also avoids overfitting the model by using a small k to capture dependency on the last kJ days. The hypothesis is that how actively one passes on the infection is affected by when they get infected. We assume that after being infected for a certain time, individuals no longer spread the infection, i.e., $\exists k$, such that $\beta_i = 0 \forall i > k$.

Also, people traveling from other regions can increase the number of infections in a given region. We assume that this infection can happen because of human mobility. Suppose $F(q, p)$ represents mobility from region q to region p. Our model is represented by the following system of equations.

$$\Delta S_t^p = -\frac{S_t^p}{N^p}\sum_{i=1}^{k}\beta_i^p \Delta I_{t-i}^p\,, \tag{1}$$

$$\Delta I_t^p = \frac{S_t^p}{N^p}\sum_{i=1}^{k}\beta_i^p (I^p_{t-(i-1)J} - I^p_{t-iJ}) + \delta\sum_q F(q,p)\frac{\sum_{i=1}^k \beta_i^q (I^q_{t-(i-1)J} - I^q_{t-iJ})}{N^q}\,. \tag{2}$$

Here, S_t^p and I_t^p represent the number of susceptible individuals and infected individuals respectively in the region p at time t. Parameter δ captures the influence of passengers coming into the region.

Note that if we set $k = 1, J = \infty$, and ignore mobility ($\delta = 0$), this reduces to Susceptible-Infected (SI) model [12]. On the other hand, with bounded $k = 1$ and $J < \infty$, the model is a variation of Suceptible-Infected-Released/Recovered (SIR) model [6], where an infected individual is active for J units of time.

3 MODELING UNREPORTED CASES

While unreported cases are not observed in the data, they affect the long term dynamics by infecting other individuals and by also reducing the number of susceptible individuals.

The individuals who are never accounted for in the reporting (in the past or the future) can be classified into two categories: (i) unreported cases - those who get infected over the course of the epidemic but do not report it; and (ii) immune/isolated cases - those who have the antibodies without being infected during the epidemic or those who are completely isolated and have 0 probability of getting infected. For unreported cases, we can add another state to our model: An individual in the i^{th} "infected" sub-state will be reported with probability γ_i^p. Thus, the total number of new reported cases is given by $\Delta R_t^p = \sum_{i=1}^k \gamma_i^p (I^p_{t-(i-1)J} - I^p_{t-iJ})$. Then the parameters will be learned by fitting the reported cases to R_t^p. The immune/isolated cases can be modeled as considering them not-susceptible, and hence not involved in the epidemic. This effectively reduces the size of the population considered for epidemic modeling. Suppose, ρ^p is the probability of a randomly selected individual in region p to be immune/isolated. Then the number of susceptible individuals at time t is given by $S_t^p = (1-\rho^p)N^p - I_t^p$, and $(1-\rho^p)N^p$ represents the reduced size of the population.

3.1 Model Simplifications for Social Distancing

In the period of social distancing, we assume that majority of the spread is "community spread" and infections due to travel across the regions (state/counties) can be ignored. For ease of notation, we drop the superscript p. For simplicity, we assume that $\gamma_i = \gamma, \forall i$. Further, we redefine I_t to be the cumulative cases that could have been reported at time t and R_t to be the cases actually reported. This allows us to ignore explicit modeling of reporting delays. Therefore, we have

$$\Delta R_t = \gamma\sum_{i=1}^{k}(I_{t-(i-1)J} - I_{t-iJ})$$
$$\text{And } R_t = \gamma I_t\,. \tag{3}$$

Combining Equation 3 with Equation 2 without the travel spread and adjusted population size, we get:

$$\frac{\Delta R_t}{\gamma} = \frac{S_t}{(1-\rho)N}\sum_{i=1}^{k}\beta_i^p\frac{(I_{t-iJ} - I_{t-(i-1)J})}{\gamma}$$
$$\implies \Delta R_t = \frac{(1-\rho)N - R_t/\gamma}{(1-\rho)N}\sum_{i=1}^{k}\beta_i^p (R_{t-iJ} - R_{t-(i-1)J})$$
$$\implies \Delta R_t = \left(1 - \frac{R_t}{\gamma(1-\rho)N}\right)\sum_{i=1}^{k}\beta_i^p (R_{t-iJ} - R_{t-(i-1)J}) \tag{4}$$

Equation 4 implies that only using the reported cases, the impact of γ and δ cannot be separately measured. Setting $\bar{\gamma} = \gamma(1-\rho) \le \gamma$, we can identify a lower bound on γ. Note that γ and ρ are not separately needed to be able to forecast the number of reported cases, and knowing $\bar{\gamma}$ is enough. However, this applies only when the infection dynamics are not changing. In the future, as the social distancing policies are relaxed, ρ is expected to change and approach 1, while γ may remain constant assuming enough testing availability. Therefore, we wish to learn γ but at this point, we can only identify $\bar{\gamma}$ which forms a lower bound for γ.

3.2 Parameter Learnability

Let $\beta = [\beta_1 \dots \beta_k]$, and $\mathbf{X_t} = [(R_t - R_{t-J}) \dots (R_{t-(k-1)J} - R_{t-kJ})]^T$. Sensitivity of ΔR with respect to γ is

$$\frac{\partial \Delta R_t}{\partial \bar{\gamma}} = \frac{R_t}{\bar{\gamma}^2 N}\mathbf{X_t}\beta. \tag{5}$$

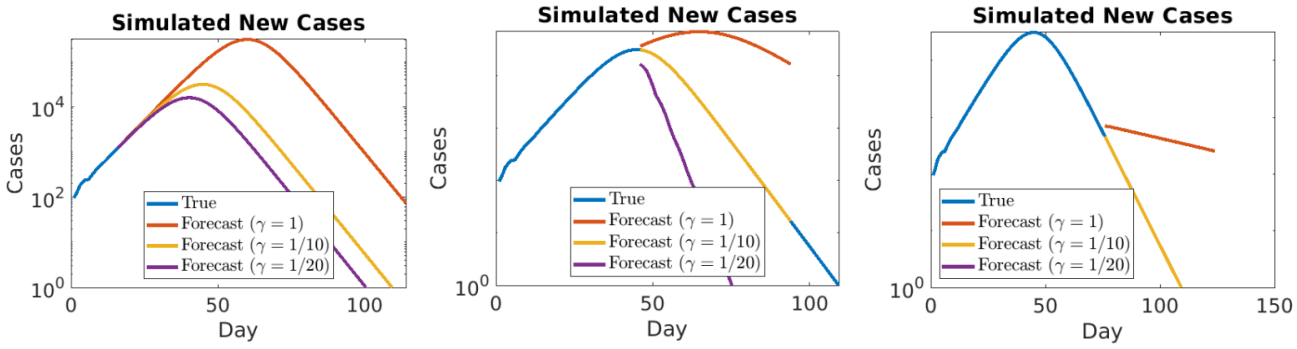
$$\frac{\partial \Delta R_t}{\partial \beta} = \left(1 - \frac{R_t}{\bar{\gamma}N}\right)\mathbf{X_t}. \tag{6}$$

In the initial phase of the epidemic, $\frac{R_{t-1}}{N} \approx 0$. Therefore, Equation 5 suggests that the number of reported cases is not sensitive to $\bar{\gamma}$ in the initial phase of the epidemic, when $\frac{R_{t-1}}{N} \approx 0$. On the other hand, Equation 6 suggests that number of new reported cases is sensitive to β.

Suppose, $\bar{\gamma}^*$ is the true value and we train by ignoring the parameter, effectively setting it to 1 to obtain β_0. Then, we get the same timeseries, if $\forall t$,

$$\left(1 - \frac{R_t}{\bar{\gamma}^* N}\right)\mathbf{X_t}\beta^* = \left(1 - \frac{R_t}{N}\right)\mathbf{X_t}\beta_0$$
$$\frac{\mathbf{X_t}\beta_0}{\mathbf{X_t}\beta^*} = 1 - \frac{R_t(1-\bar{\gamma})}{\bar{\gamma}(N - R_t)}\,, \tag{7}$$

which is close to 1, when $R_t \ll N$. Figure 3 demonstrates this fact. We simulate an epidemic with $\beta = [0.4\ 0.2], N = 1,000,000$ and $\bar{\gamma} = \gamma = 1/10$. We then attempt to "forecast" assuming the knowledge of β, and various values of $\bar{\gamma} = \gamma = 1, 1/10$ and $1/20$. Observe that in the initial phase of the epidemic (Figure 2a) all three trends are similar until they get close to the peak. Starting at the peak (Figure 2b) and after the peak (Figure 2c), with the same initial values and β, significantly different forecasts are obtained by varying γ. By setting $k = 1$ in Equation 7, the following can be easily proved.



(a) Initial Phase **(b) Around Peak** **(c) 'Tail' Phase**

Figure 2: Effect of varying γ at different phases of the epidemic on the reported cases.

Theorem 3.1. *For a given R_f, $\exists \epsilon > 0$, such that $\forall R_t \leq R_f$, $0 \leq \frac{\beta^* - \beta_0}{\beta^*} \leq \epsilon$.*

Proof. From Equation 7, easy to see that $\beta_0 \leq \beta^*$. Setting $\epsilon = \frac{R_f(1-\bar{\gamma}^*)}{\bar{\gamma}(N-R_f)}$ completes the proof. □

Theorem 3.1 suggests that early part of the epidemic is not reliable for learning $\bar{\gamma}$. However, this does not imply that we should always prefer a high value of t in the following where we explore the effect of the "tail" part of the epidemic on the learnability of $\bar{\gamma}$.

Lemma 3.2. *There exists τ such that $\bar{\gamma}$ that describes the data for $R_t > R_\tau$ is not unique.*

Proof. We prove this by showing that there is a t_u such that for $t > t_u$, there are at least two sets of parameter $(\beta_1, \bar{\gamma}_1)$ and $(\beta_2, \bar{\gamma}_2)$ that fit the data for $t > t_u$, i.e., the following has a feasible solution.

$$\Delta R_t = \left(1 - \frac{R_t}{\bar{\gamma}_1 N}\right) \beta_1 \mathbf{\Delta X_t} = \left(1 - \frac{R_t}{\bar{\gamma}_2 N}\right) \beta_2 \mathbf{\Delta X_t}.$$

Setting $k = 1$, $\mathbf{\Delta X_t}$ becomes a scalar. After some algebraic manipulations, we get

$$\bar{\gamma}_2 = \frac{(\beta_2/\beta_1)\bar{\gamma}_1 R_t}{R_t - (1 - \beta_2/\beta_1)\bar{\gamma}_1 N} \tag{8}$$

This is a valid solution, if $0 < \bar{\gamma}_2 \leq 1$. Without loss of generality, we can assume $\beta_2 < \beta_1$. Then

$$\bar{\gamma}_2 > 0 \implies R_t > \bar{\gamma}_1 N(1 - (\beta_2/\beta_1)),$$
$$\text{And } \bar{\gamma}_2 \leq 1 \implies R_t > \bar{\gamma}_1 \frac{N(1 - (\beta_2/\beta_1))}{1 - (\beta_2/\beta_1)\bar{\gamma}_1}.$$

Therefore, if the data contains R_t such that the above holds for all t, then at least two solutions for $(\beta, \bar{\gamma})$ exist. □

The above lemma suggests that we should not attempt to learn the parameters solely from the "tail" of the epidemic. However, using the beginning part only, we cannot reliably learn $\bar{\gamma}$ as discussed earlier. Next, we identify what data needs to be included to guarantee accurate learning of $\bar{\gamma}$.

Theorem 3.3. *Suppose, $(\beta_0, \bar{\gamma}_0)$ is a solution obtained from the given data. Let $\beta^* \geq \beta_0 \geq (1-\epsilon)\beta^*$, for some $0 \leq \epsilon < 1$. Then for any R_τ, there exists a $0 < \delta < 1$ such that choosing data $R_t > R_\tau$ guarantees that $(1-\delta)\bar{\gamma}_0 \leq \bar{\gamma}^* \leq \bar{\gamma}_0$.*

Proof. Since, $\beta_0 \leq \beta^*$, $\bar{\gamma}_0 \geq \bar{\gamma}^*$. Suppose, for some $\delta > 0$, we wish to prove that that $\bar{\gamma}^* \geq (1-\delta)\bar{\gamma}_0$. Assume the contrary that $\bar{\gamma}^* < (1-\delta)\bar{\gamma}_0$. Then, using Equation 8,

$$\bar{\gamma}_0 = \frac{\bar{\gamma}^*(\beta_0/\beta^*)R_t/N}{R_t/N - (1 - \beta_0/\beta^*)\bar{\gamma}^*}$$
$$\implies \frac{\bar{\gamma}^*}{1-\delta} < \frac{\bar{\gamma}^*(\beta_0/\beta^*)R_t/N}{R_t/N - (1 - \beta_0/\beta^*)\bar{\gamma}^*}$$

Using $\beta^* \geq \beta_0$ in the numerator and $\beta_0 \geq (1-\epsilon)\beta^*$ in the denominator of the RHS, we get

$$\frac{1}{1-\delta} < \frac{R_t/N}{R_t/N - \epsilon\bar{\gamma}^*}$$
$$\implies R_t < \frac{N\epsilon\bar{\gamma}^*}{\delta} \leq \frac{N\epsilon\bar{\gamma}_0}{\delta}.$$

Therefore, if $R_t \geq R_\tau = \frac{N\epsilon\bar{\gamma}_0}{\delta}$ the above is not feasible, and thus $\bar{\gamma}^* \geq (1-\delta)\bar{\gamma}_0$. □

Finally, we present how $\bar{\gamma}$ affects the peak of the epidemic.

Theorem 3.4. *If the peak of new cases happens when the total cases are R_{peak}, then*

$$\bar{\gamma} \approx \frac{R_{peak}/N}{1 - \frac{1}{J\|\beta\|_1}}, \tag{9}$$

where $\|\beta\|_1 = \sum_i \beta_i$.

Proof. At the peak, we assume that ΔR_t remains constant for a window of $kJ + 1$ time steps, i.e., $\Delta R_t = r, \forall t = \tau, \tau - 1, \ldots, \tau - kJ$. Then $\beta \mathbf{X} = J\|\beta\|_1$. Therefore, we have

$$r \approx \left(1 - \frac{R_\tau}{\bar{\gamma}N}\right) J\|\beta\|_1 r$$
$$\implies \bar{\gamma} \approx \frac{R_\tau/N}{1 - \frac{1}{J\|\beta\|_1}}. \tag{10}$$

□

Next, we utilize Theorems 3.1, 3.3, and 3.4 to learn the parameters β and $\bar{\gamma}$.

4 LEARNING

Unlike [11] where the goal was to perform forecasts in an adaptive fashion even during changing policies, here, our main goal is identifying $\bar{\gamma}$. This knowledge can then be used for performing forecasts. For learning, we first manually identify and remove the part of the data where the effect of social distancing is visible. For instance, in Figure 1 the initial part shows rapid rise when no precautions were taken. This step is necessary for our axiom that the remaining data can be assumed to follow the same dynamics, i.e, has a true unique $(\beta, \bar{\gamma})$.

4.1 Fixed Infection Rate Method

In this approach we utilize the fact that the effect of the unreported cases is not seen in the initial part of the infection. Therefore, we consider an initial part of the reported cases data up to time t_f. We use this initial part to train the model to learn β_0 by fixing $\bar{\gamma} = 1$. Then, by Theorem 3.1, $\beta^* \leq \beta_0 \leq (1-\epsilon)\beta^*$, for some ϵ.

Then, we train a linear model by fixing the previously learned β_0 as a constant and learn $\bar{\gamma}_0$. We identify the largest value of $R_t = R_{max}$ available in the dataset. From Theorem 3.3, it follows that setting

$$\delta = \frac{N\epsilon\bar{\gamma}_0}{R_{max}} \tag{11}$$

ensures that there is at least one data point for the model to identify $\bar{\gamma}_0$ such that $(1-\delta)\bar{\gamma}_0 \leq \bar{\gamma}^*$.

Next we discuss, how to identify the value of ϵ. Note that ϵ as calculated in Theorem 3.1 relies on $\bar{\gamma}^*$, which is not known. We use the fact that a δ must exist such that $\bar{\gamma}^* \geq (1-\delta)\bar{\gamma}_0$. Using this bound in Theorem 3.1, there exists a δ for which

$$\epsilon = \frac{R_f\left(1-(1-\delta)\bar{\gamma}_0\right)}{(N-R_f)(1-\delta)\bar{\gamma}_0}. \tag{12}$$

Putting the value ϵ in Equation 11 results in a quadratic equation in δ with the smaller root

$$\delta = \frac{1-\frac{R_f}{N}-\bar{\gamma}_0 - \sqrt{\left(1-\frac{R_f}{N}-\bar{\gamma}_0\right)^2 - 4\left(1-\frac{R_f}{N}\right)\left(\frac{R_f}{R_{max}}-\bar{\gamma}_0\right)}}{2\left(1-\frac{R_f}{N}\right)} \tag{13}$$

- Test1 (hard): Is δ a real number and in $(0,1)$? If not, then the method fails, as we are unable to guarantee reliability.
- Test2 (soft): For a given $\delta_3 > 0$, and the number of cumulative reported cases at the peak R_{peak}, $(1-\delta_3)\bar{\gamma}_0 \leq \frac{R_{peak}/N}{1-1/(J\|\beta_0\|)} \leq (1+\delta_3)\bar{\gamma}_0$? This is a "soft" test in the sense that it is based on an approximation and can be performed only if the "peak" is available. Identifying the actual peak is difficult due to noisy data, and thus δ_3 provides a soft margin for the peak.

The parameters are learned using least square estimation:

$$LSE = \sum_{t=\tau}^{T}\left(\left(1-\frac{\hat{R}_t}{\bar{\gamma}N}\right)\mathbf{X_t}\beta - \Delta\hat{R}_t\right)^2 \tag{14}$$

$$\tag{15}$$

Here $\hat{R}_t \forall t$ are true observed values. Least square optimization is performed using trust-region reflective algorithm [7]. Note that the above approach may be prone to noisy initial values. However, we smooth the data before learning the parameters to avoid noise.

Alternatively, the initial values $\Delta R_{\tau-J}, \ldots, \Delta R_{\tau-1}$ can also be treated as learnable parameters. In this case, we fit the curve obtained by the recurrence relation $\Delta R_t = \left(1-\frac{R_t}{\bar{\gamma}N}\right)X_t\beta$ to the observed data $< \Delta R_{\tau-J}, \ldots \Delta R_{\tau-1}, R_\tau, R_{\tau+1} \ldots \Delta R_T >$. While this approach is better for dealing with noisy data, it may be prone to overfitting due to additional J parameters. Least square optimization is performed using trust-region reflective algorithm [7].

4.2 Heuristic Methods

We also propose treating β and $\bar{\gamma}$ simultaneously as learnable parameters as a heuristic approach. Since, Theorems 3.1 and 3.3 do not apply, we cannot perform Test1 to ensure reliability. However, we can perform Test2. As in the case of Fixed Infection Rate Method, we have two ways of learning the heuristic models.

Non-linear Incremental Learning. The parameters are learned using least square estimation:

$$LSE = \sum_{t=\tau}^{T}\left(\left(1-\frac{\hat{R}_t}{\bar{\gamma}N}\right)\mathbf{X_t}\beta - \Delta\hat{R}_t\right)^2 \tag{16}$$

$$\tag{17}$$

Non-linear Curve Fitting. Learning is performed by fitting a curve over time as opposed to a linear model by treating the initial values $\Delta R_{\tau-J}, \ldots, \Delta R_{\tau-1}$ as learnable parameters as well.

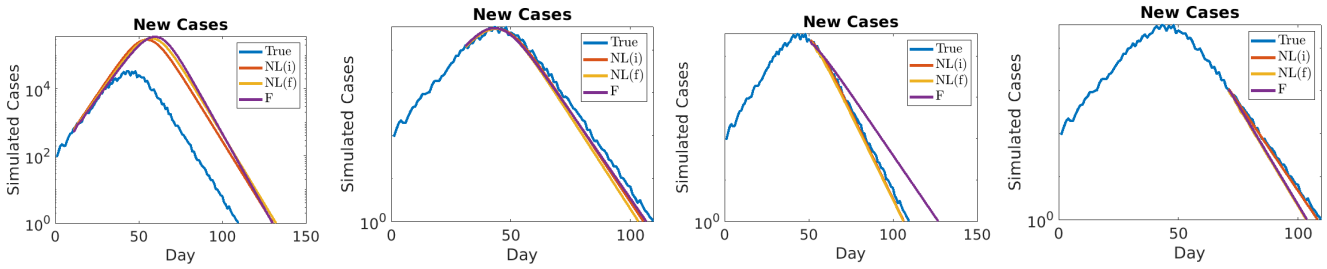
It is possible to derive reliability bounds on these heuristics as well, however, they are unlikely to be useful. We wish to identify a lower bound on reporting probability, therefore, if $\bar{\gamma}_0 \leq \bar{\gamma}^*$, then we have nothing to prove. Suppose, $\bar{\gamma}_0 \geq \bar{\gamma}^*$. Then we would like to show that $\bar{\gamma}^* \leq (1-\delta)\bar{\gamma}_0$, which follows the same derivation as Equation 11. Choosing an epsilon here is difficult – Using the scheme as in Fixed Infection Rate algorithm leads to $\delta > 1$. A valid choice is $\epsilon = 1-\beta_0$ (obtained using $\beta^* \leq 1$), which would result in $\delta = N(1-\beta_0)\bar{\gamma}_0./R_{max}$. This is often larger than 1 in practice, and thus not useful.

Here, we have chosen $k = 1$ as our Test1 is derived for scalar β. However, the above algorithms can be used (without reliability tests) for any value of k with Test2. In Section 5.2 we have explored the effectiveness of the above algorithms for $k > 1$.

5 EXPERIMENTS

5.1 Setup

We obtained all the reported cases fom JHU CSSE COVID19 dataset [1]. Particularly we extracted county level data for New York City and Los Angeles. These were used because these two counties have performed serology tests with initial estimation of number of unreported cases. We further performed experiments on all US states, most of which did not pass our tests for reliability. Here we will report the results on New York, Illinois, Massachusetts, and New Jersey - four of the states with the most reported cases. Population of the counties and states were obtained from the US Census Bureau [4].



(a) (10, 30) (b) (30, 50) (c) (50, 70) (d) (70, 90)

Figure 3: Fitting the models over various intervals in the simulated data.

The county data showed significant amount of noise, and so, it was smoothed with moving average over two weeks, before applying our learning algorithms. The state-level timeseries were relatively less noisy, and were smoothed with moving average over one week. All the code was written in MATLAB and is available online[1]. We set $k = 1$ and $J = 7$ for the US counties and states. The choice for J was driven by observed weekly periodicity in the data [**?**].

5.2 Simulation

To demonstrate the effectiveness and limitations of the three approaches, we used the same setting as in Section 3.2 but with added noise to simulate an epidemic. We attempted to retrieve the parameters $(\beta, \bar{\gamma})$ using our three learning approaches - Non-linear Incremental Learning NL(i), Non-linear Curve Fitting NL(f), and Fixed Infection Rate Learning (F). These methods learn the models fitted on data for $T \in (\tau_1, \tau_2)$ for various intervals. Here, Fixed Infection Rate approach is simplified to use use $(1, \tau_1)$ to first identify β, and (τ_1, τ_2) to identify $\bar{\gamma}$, without any reliability guarantee. Note that here, $k = 2$ and our reliability analysis applies only to $k = 1$. Regardless, we perform these experiments to observe the effect of $k > 1$. Figure 3 shows the fit along with forecasts until the end of the epidemic. Observe that for the interval $(30, 50)$ all methods are able to accurately forecast. 'NL(i)' and 'NL(f)' are able to accurately forecast for the interval $(50, 70)$. It also seems that the three methods accurately forecast by learning on the interval $(70, 90)$. To assess whether these fits actually retrieve the values of $\bar{\gamma}$, we present the learned parameters in Table 1. Note that only for the interval $(30, 50)$ all three methods are able to identify $\bar{\gamma}$ close to its original value, i.e., 0.1. While we were able to see accurate forecasts for the interval $(70, 90)$, the learned values of $\bar{\gamma}$ are far from the true value. This reinforces our claim that there exists a certain window of data which is needed to accurately learn $\bar{\gamma}$.

5.3 Results: US Counties

Figure 4 shows the model fit obtained on New York City and Los Angeles. Recall that $\bar{\gamma} = (1 - \rho)\gamma$, where γ is the probability of reporting an infected case. Therefore, $1/\bar{\gamma}$ forms the upper bound on the estimated number of total cases as a factor of reported cases. We report these upper bounds in Table 2. We have shown the factors obtained using 95% confidence interval on $\bar{\gamma}$. Additionally, for Fixed Infection Rate learning, we have provided an additional bound obtained from Theorem 3.3 with δ obtained from Equation 13. The three methods result in factors close to each other (39-44) for New York City. However, the reliable bound obtained was 63. Figure 4 suggests that all three methods produce good fit for New York City. Note that this factor provides an upper bound on the actual ratio of total to reported cases. As an illustration, if we agree that the bound obtained for NYC is 40 and $\rho = 0.5$, i.e., half of the population was able to completely isolate itself reducing its probability of infection to zero, then the the number of true cases will 0.5×36, i.e, 18 times of the reported cases. On the other hand, none of the results for Los Angeles were sensible (see Table 2). 'OOR' indicates that the 95% confidence interval was out of the feasible range of the solution. For the method 'F', Test1 failed. It implies that it may be too early to reliably estimate the upper bound of this factor from Los Angeles data.

Note that antibody tests in New York in April estimated that 24.7% of the entire population were infected[2]. Based on the population of New York City and the number of reported cases at the time, this translates to actual cases being roughly 13.8 times the reported cases.

5.4 Results: US States

We also estimated the bound on the total number of actual cases as a factor of reported cases for various states. Table 3 shows the results for New York, Illinois, Massachusetts, and New Jersey. All the states presented here, passed Test1 and Test2. Figure 5 shows the model fit obtained using the learned parameters.

For New York our methods estimated that the bound on total cases is 23-25 times of the reported cases with the reliable worst case bound being 35.17. Note that the state-wide antibodies study in early May estimated that 12.3% of the state population was infected[3]. This translates to actual cases being roughly 7.6 times the reported cases. For Illinois, Massachusetts and New Jersey, this factor is roughly 33-37, 27-31, and 19-21, with worst case upper bound being 40.86, 38.28, and 29.22, respectively. If we assume that these states are similar enough that they have the same probability γ of reporting and the same fraction of population that is completely isolated,

[1] https://github.com/scc-usc/ReCOVER-COVID-19

[2] https://www.livescience.com/covid-antibody-test-results-new-york-test.html

[3] https://www.governor.ny.gov/news/amid-ongoing-covid-19-pandemic-governor-cuomo-announces-results-completed-antibody-testing

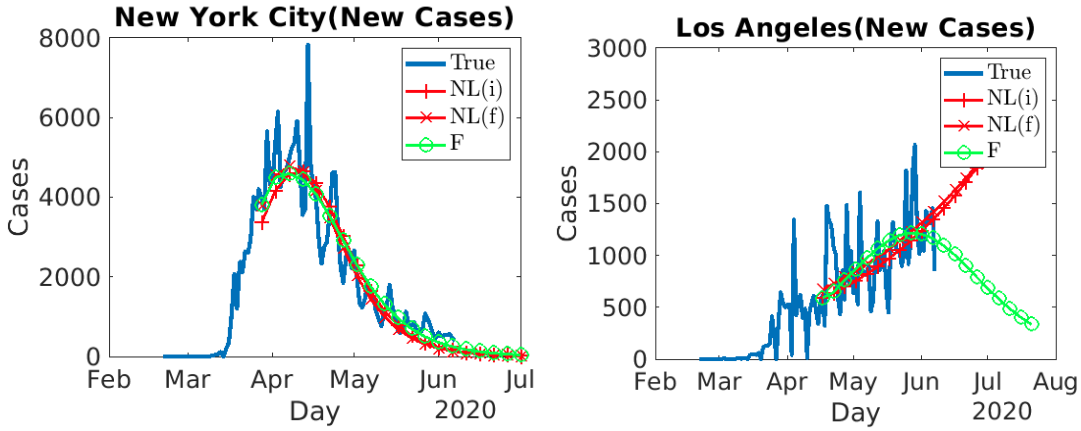


Figure 4: Model fittings for counties using our three algorithms.

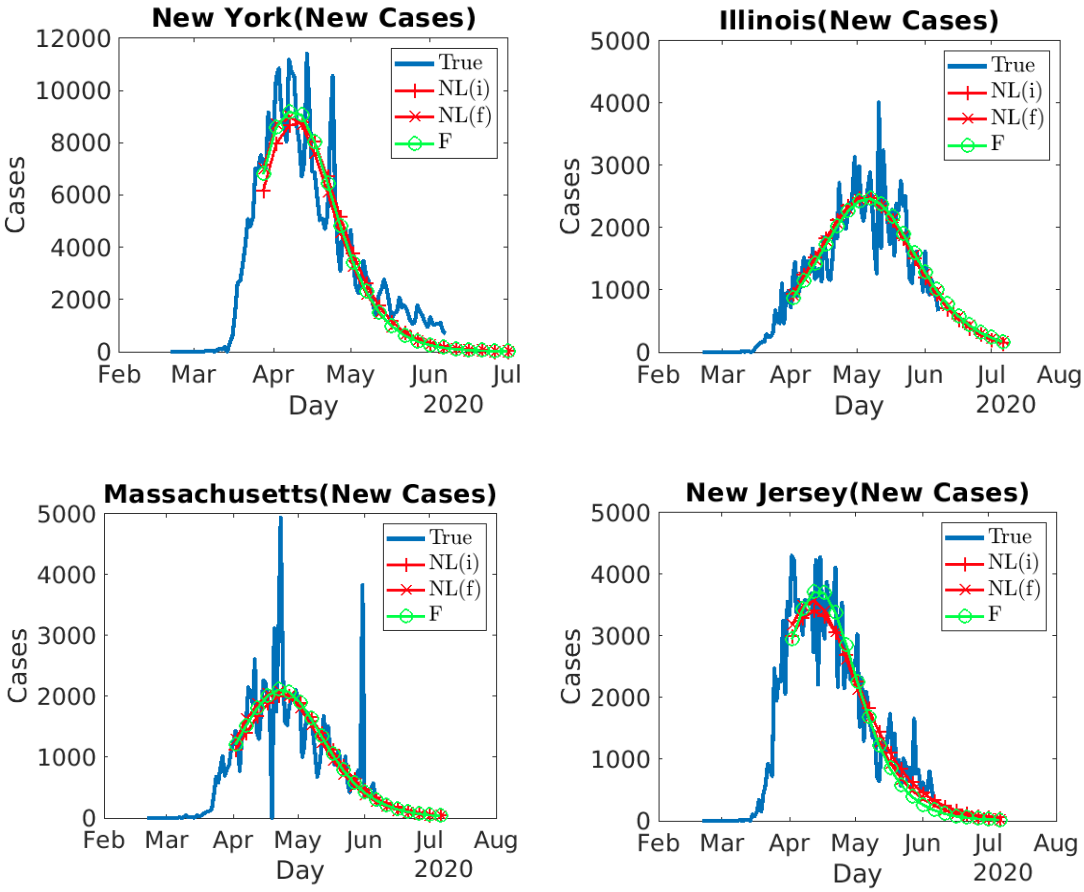


Figure 5: Model fittings for states using our three algorithms.

Table 1: Learned parameters (β_1, β_2), $\bar{\gamma}$ from simulated experiments. The true value of $\bar{\gamma} = 0.1$.

(τ_1, τ_2)	NL(i)	NL(f)	F
(10, 30)	(0.1723, 0.3619), 1	(0.3487, 0.02453), 1	(0.5569, 0.1071), 1
(30, 50)	(0.4408, 0.1793, **0.934**	(0.4408, 0.1793), **0.092**	(0.1750, 0.3620), **0.1095**
(50, 70)	(0.2064, 0.4099), **0.1036**	(0.5153, 0.0991), **0.0916**	(0, 0.3652), 0.2440
(70, 90)	(0.0898, 0.0853), 1	(0.1246, 0), 0.7796	(0, 0.3438), 0.1956

Table 2: Estimated upper bound on number of total cases as a factor of reported cases for the counties.

States	NL(i)	NL(j)	F
New York City	41.2712 - 44.2499 - 47.6919	40.2654 - 42.8327 - 45.7496	38.8609 - 39.7612 - 62.9513
Los Angeles	OOR	OOR	(x)Test1

Table 3: Estimated upper bound on number of total cases as a factor of reported cases.

States	NL(i)	NL(j)	F
New York	22.2565 - **24.8175** - 28.0445	21.1909 - **23.0848** - 25.3503	22.7918 - **23.4891** - 35.1732
Illinois	30.3602 -**33.4322** - 37.1959	31.9536 - **33.4813** - 35.1624	36.2494 - **36.7681** - 40.8638
Massachusetts	24.1917 - **27.3885** - 31.5589	25.4437 - **27.5206** - 29.9668	30.3027 - **31.4906** - 38.2838
New Jersey	18.5704 - **20.3788** - 22.5774	19.0749 - **20.3332** - 21.7692	18.6939 - **19.0698** - 29.2222

then we can conclude that for all these states, the true cases cannot be more than 29.22 times, which satisfies all the upper bounds. All four states passed Test2 (peak test) with $\delta_3 = 0.2$.

We have not presented results for the US at country-level due to high heterogeneity in the infection trends of various states. Therefore, learning a single parameter for the entire country may not be accurate, and it may be better to learn separately for different states.

6 CONCLUSIONS

We have proposed Fixed Infection Rate algorithm to reliably estimate a bound on number of unreported cases. The algorithm is built upon key theorems that identify limitations of learnability of reporting probability. We have also proposed two heuristics that learn this bound but do not provide guarantees. We demonstrate through simulated experiments that all three methods are able to identify the bound correctly on certain regions of the epidemic. We emphasize that these algorithms learn $\bar{\gamma}$ which combines the effect of reporting probability and isolated population. Particularly, if a fraction ρ of the total population completely isolates itself, thus getting removed from the epidemic, then $\bar{\gamma} = (1 - \rho)\gamma$, where γ is the probability of reporting a case (symptomatic or asymptomatic). Hence, $\bar{\gamma}$ forms the lower limit for reporting probability. In other words we can find an upper bound on total number of infected cases. Applying our algorithm on the data during the social distancing phase, we conclude with high confidence that the actual number of cases cannot be more than 35 times in New York, 40 times in Illinois, 38 times in Massachusetts, and 29 times in New Jersey, than the reported cases. In future work, we will explore obtaining tighter bounds, when the precautions are relaxed and the fraction of isolated population ρ is reduced. We will further explore how to utilize data across changing dynamics due to changing policies to strengthen these bounds.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation Award No. 2027007.

REFERENCES

[1] [n.d.]. 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE. https://github.com/CSSEGISandData/COVID-19.
[2] [n.d.]. CHIKV Challenge Announces Winners, Progress toward Forecasting the Spread of Infectious Diseases. https://www.darpa.mil/news-events/2015-05-27.
[3] [n.d.]. DARPA forecasting chikungunya challenge. https://www.innocentive.com/ar/challenge/9933617.
[4] [n.d.]. State Population Totals: 2010-2019. https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html.
[5] Eran Bendavid, Bianca Mulaney, Neeraj Sood, Soleil Shah, Emilia Ling, Rebecca Bromley-Dulfano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra, James Tedrow, et al. 2020. COVID-19 Antibody Seroprevalence in Santa Clara County, California. *MedRxiv* (2020).
[6] Ottar N Bjørnstad, Bärbel F Finkenstädt, and Bryan T Grenfell. 2002. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecological monographs* 72, 2 (2002), 169–184.
[7] Thomas F Coleman and Yuying Li. 1996. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on optimization* 6, 2 (1996), 418–445.
[8] Arnaud Ducrot, P Magal, Thanh Nguyen, and GF Webb. 2020. Identifying the number of unreported cases in SIR epidemic models. *Mathematical medicine and biology: a journal of the IMA* 37, 2 (2020), 243–261.
[9] Zhihua Liu, Pierre Magal, Ousmane Seydi, and Glenn Webb. 2020. Understanding unreported cases in the COVID-19 epidemic outbreak in Wuhan, China, and the importance of major public health interventions. *Biology* 9, 3 (2020), 50.
[10] Pierre Magal and Glenn Webb. 2018. The parameter identification problem for SIR epidemic models: identifying unreported cases. *Journal of mathematical biology* 77, 6-7 (2018), 1629–1648.
[11] Ajitesh Srivastava and Viktor K Prasanna. 2020. Learning to Forecast and Forecasting to Learn from the COVID-19 Pandemic. *arXiv preprint arXiv:2004.11372* (2020).
[12] Tao Zhou, Jian-Guo Liu, Wen-Jie Bai, Guanrong Chen, and Bing-Hong Wang. 2006. Behaviors of susceptible-infected epidemics on scale-free networks with identical infectivity. *Physical Review E* 74, 5 (2006), 056109.