Machine Learning the Phenomenology of COVID-19 From Early Infection Dynamics

Malik Magdon-Ismail
magdon@cs.rpi.edu
Computer Science Department
Rensselaer Ploytechnic Institute
110 8th Street, Troy, NY 12180, USA

March 17, 2020

Abstract

We present a data-driven machine learning analysis of COVID-19 from its *early* infection dynamics, with the goal of extracting actionable public health insights. We focus on the transmission dynamics in the USA starting from the first confirmed infection on January 21 2020. We find that COVID-19 has a strong infectious force if left unchecked, with a doubling time of under 3 days. However it is not particularly virulent. Our methods may be of general interest.

1 Introduction

On March 1 on ABC's This Week, US Secretary of Health and Human Services Alex Azar told host George Stephanopoulos that asymptomatic spread of COVID-19 is "not the major driver" of the spread of the new coronavirus, a sentiment corroborated on the CDC's website. Meanwhile several other researchers disagreed, for example Dr. Osterholm, director of the Center for Infectious Disease Research and Policy at the University of Minnesota and Dr. Schaffner of the Vanderbilt University School of Medicine and longtime adviser to the CDC (see the CNN article, Cohen (2020)). At the time, there was clearly no unimodal opinion. What does the data say? In a recent article, Chinazzi *et al.* (2020) used the Global Epidemic and Mobility Model (GLEAM) to perform infection scenario analyses on China COVID-19 data, using a networked meta-population model focused on transportation hubs. A similar model for the US was reported in Wilson (2020) where their web-app predicts from 150,000 to 1.4 million infected cases by April 30, depending on the intervention scenario. Such complex scenario analysis requires user input such as infection sites and properties of the contagion, and the model is taken at face value. A majority of infection sites are hidden, especially if asymptomatic transmission is significant, and a range of contagion parameters must be deduced, perhaps using domain expertise.

There is a clear need for generating actionable intelligence on the nature of a pandemic from *early* and minimal data. Data driven approaches are powerful. A long range regression analysis of COVID-19 out to 2025 on US data using immunological, epidemiological and seasonal effects is given in Kissler *et al.* (2020), which predicts recurrent outbreaks. We also follow a pure data-driven machine learning approach to understand early dynamics of COVID-19 on the first 54 days of US confirmed infection reports (downloadable from the European Center For Disease Control). We address the challenge of real-time data-intelligence from early data. Our data-approach is simple, requires minimal data or human input and generates fast actionable insights. Is asymptomatic

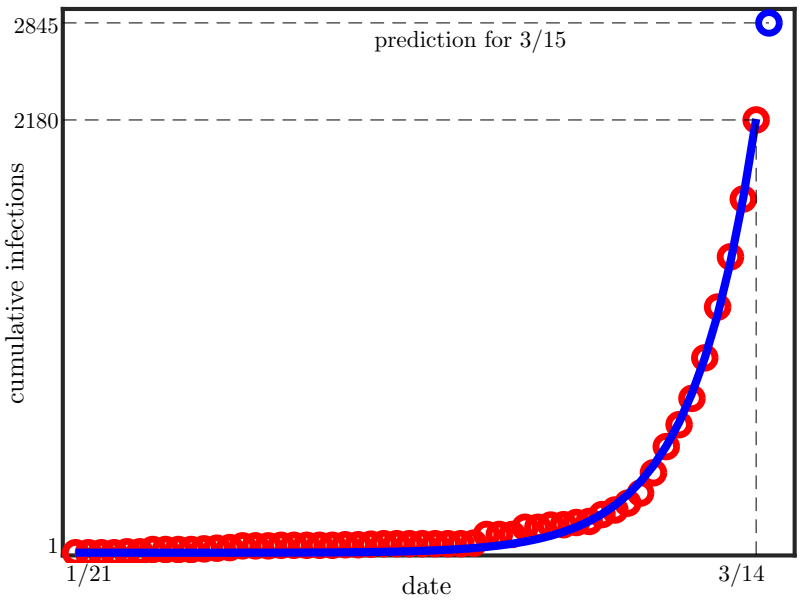

Figure 1: Early dynamics of (USA) COVID-19 and model prediction.

spread significant? Our data-driven analysis says yes, emphatically. What are the properties of the contagion, for example is it unusually virulent? Our analysis says no.

Early data is both a curse and a blessing. The curse is that early data means you do not have much information, so quantitative models must be simple and robust if they are to be learned from the data. The blessing is that early data is a peek at the pandemic as it roams free, unchecked by public health protocols, for to learn the true intentions of the lion, you must follow the beast on the savanna, not in the zoo. As we will see, much can be learned from this early data and these insights early in the game, that are not based on opinion but data, are crucial to public health governance.

We analyze confirmed COVID-19 cases as they arrive day by day from January 21 to March 14. We test the predictions on March 15 and 16 and we eagerly await March 17 data. Our model predictions up to March 15 are in Figure 1. Red circles are cumulative confirmed infections, the blue curve is our learned model and the blue circle is our prediction for March 15, 665 new confirmed infections. Qualitatively the model captures the data. The blue model curve is, in some sense, an optimal fit to the data, with an *asymptomatic* infectious force if left unchecked of 30% new cases per day (doubling in about 2.6 days) and a virulence of $\gamma = 0.14\%$ (1 or 2 in a thousand conversions from mild infection to serious). Such intelligence extracted from the learned model has public health significance. The spread rate is high, hence protocols should be invoked to curtail that spread, to prevent overloading of the health-system. On the other hand, we should not panic, the virulence is not astonishingly high.

2 Model, Data and Analysis

Our goal is a simple robust model. By far, the majority of disease models are based on aggregation of individuals according to disease status, such as SI, SIR, SIS Kermack and McKendrick (1927); Bailey (1957); Anderson and May (1992). We use a similar model by considering a mild infected state which can transition to a serious state. Early data allows us to make simplifying assumptions. In the early phase, when public health protocols have not kicked in, a confirmed infection is self-

reported. That is, *you* decide to get tested. Why? Because the condition became serious. That is an important bit of information. A confirmed case is a transition from mild infection to serious. This is not true later when, for example, public health protocols may institute randomized testing. Let us assume that at time t there are $C(t)$ confirmed cases and correspondingly $M(t)$ mild unreported (asymptomatic) infections. The new confirmed cases at time t correspond to mild infections at some earlier time $t-k$ which have now transitioned to serious and hence got self-reported. Let us suppose that a fraction γ of those mild cases transitioned to serious, $C(t) = C(t-1) + \gamma M(t-k)$. Another advantage of early dynamics is that we may approximate the growth from each infection as independent and exponential, according to the infectious force of the disease. So $M(t) = \beta M(t-1) - \gamma M(t-k) + \alpha C(t-1)$. Here, the second term is the loss of mild cases that transitioned to serious and confirmed, and the third term accounts for new infections from confirmed cases. We will further simplify and assume that confirmed cases are fully quarantined and so $\alpha = 0$. Our simplified model is:

$$\begin{aligned} S(t) &= S(t-1) + \gamma M(t-k) & S(t) = 1 \text{ for } 0 \le t < k \\ M(t) &= \beta M(t-1) - \gamma M(t-k) & M(0) = M_0. \end{aligned} \tag{1}$$

We set $t = 0$ at the first confirmed infection (Jan 21 in USA). There are four parameters.

β, the (asymptomatic) infectious force parameter governing the exponential spreading
γ, the virulence, the fraction of mild cases that become serious some time later
k, the lag time for mild (asymptomatic) infection to become serious (incubation time)
M_0, The number of unconfirmed (mild) infections at time 0

We may assess the quality of given set of parameters β, γ, k, M_0 by how well they reproduce the observed dynamics in Figure 1, using an error-measure, see for example Abu-Mostafa *et al.* (2012). We use a combination of the root-mean-squared-error and the root-mean-squared-percentage-error between the observed dynamics and the model predictions. The model parameters which optimally fit the data are

$$\beta^* = 1.30 \qquad \gamma^* = 0.0014 \qquad k^* = 10 \text{ days} \qquad M_0^* = 4 \tag{2}$$

The (asymptomatic) infectious force, β, is very high, and corresponds to a doubling time of 2.6 days. On the other hand, the virulence at 0.14% is small, comparable to a standard flu. That is promising. The incubation period of 10 days seems in line with physician observations. The data analysis predicts that when the first confirmed case appeared, there were 4 other infections in the USA. The parameters β^*, γ^* and M_0^* are new knowledge, gained with relative ease by calibrating a simple robust model to the early dynamics. But, these optimal parameters are not the whole story, when it comes to prediction.

To understand the model fitting better, let us walk through the process of obtaining the optimal parameters $(\beta^*, \gamma^*, k^*, M_0^*)$. First we determine the incubation period k and the initial number of unseen mild infections M_0. To do so, we exhaustively search a grid over (k, M_0) and for each pair of values, we perform a gradient descent with 4 million iterations to identify the optimal β and γ for that specific k, M_0. We evaluate the best achievable fits for various (k, M_0) in Figure 2(a). The deep-blue region corresponds to essentially equivalent models within 0.5% of the optimal fit, our (user defined) error tolerance. The deep-blue region shows the extent to which the model is ill-identified by the data. Indeed, all these deep-blue models equally fit the data which results in a range of predictions. For robustness, we pick the white square in the middle of the deep-blue region, but note here that it is only one of the models which are consistent with the data. In making predictions, we should consider all equivalent models to get a range for the predictions that are

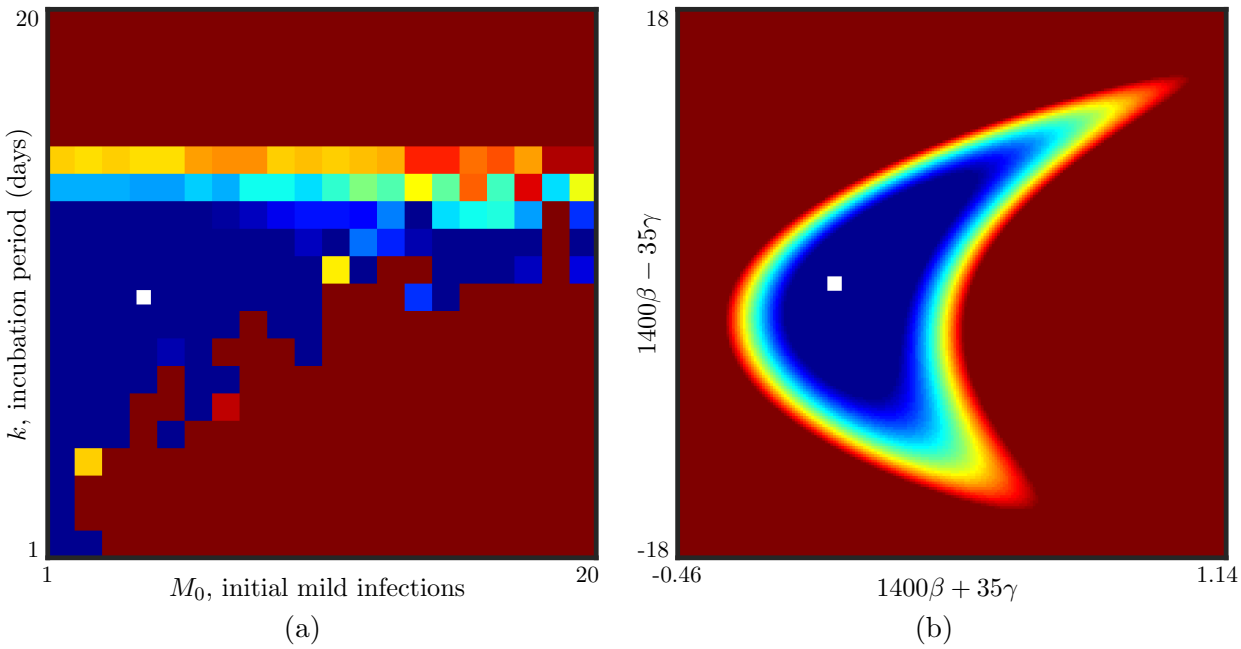

Figure 2: Uncertainty in model estimation. (a) Optimal fit-error over (β, γ) for different choices of incubation period and initial infections. Blue is better fit, red is worse fit. The deep-blue region corresponds to comparable models, within 0.5% of the optimal fit. The white square is the model chosen, $(k^*, M_0^*) = (10, 4)$ which has optimal fit and also is "robust" being in the middle of the deep-blue regions. (b) Model fit for the chosen k^* and M_0^*. Again, blue is a better fit and the deep-blue region is an equivalence class of optimal models. We choose the "robust" model corresponding to the white square in the middle of the deep-blue region, $(\beta^*, \gamma^*) = (1.30, 0.0014)$ The deep-blue regions represent uncertainty.

all equally consistent with the data. Similarly, in Figure 2(b), we fix k^* and M_0^* to their optimal robust values and show the uncertainty with respect to β and γ (the deep-blue region). Again, we pick the white square in the middle of the deep-blue region of equivalent models with respect to the data. Hence we arrive at our optimal parameters in Equation (2).

3 Predictions and Discussion

As mentioned in the previous section, to get honest estimates, we must consider all models which are equally consistent with the data, indicated by the deep-blue regions in Figure 2. We arrive at the following ranges for our estimates, based on a 0.5% error tolerance.

$\beta(\beta^* = 1.30)$	$\gamma(\gamma^* = 0.14\%)$	$k(k^* = 10$ days$)$	$M_0(M_0^* = 4)$	Asymptomatic on 03/14
$[1.3, 1.31]$	$[0.03\%, 1.2\%]$	[1 day, 13 days]	$[1, 12]$	[1.3 million, 16 million]

The model says the (asymptomatic) infectious force is very strong, around 30%, however it is not that virulent, converting at most 1.2% of mild infections to serious, comparable to a flu. There is significant uncertainty in the incubation period, from 1 up to 13 days, and there could have been up to 12 unknown initial infections when first case was confirmed. As of March 14, the model says there are on the order of a million unknown infections.

We now come to the predictions of new infections, how the model predicts $S(t)$ will evolve beyond March 14, see Table 1. We could have added March 15 and 16 to the data and re-calibrated, but then we wouldn't have a test prediction to validate on. You may complain about the large uncertainty-range for each prediction. Unfortunately, that is what the data tells us. There is

March	15	16	17	18	19
New infections	657	866	1130	1475	1924
Range	[657,2861]	[856,3732]	[1114,4867]	[1450,6348]	[1888,8280]
Observed	**777**	**823**	???	???	???

March	20	21	22	23	24
New infections	2510	3275	4272	5574	7271
Range	[2459,10800]	[3201,14088]	[4168,18375]	[5427,23968]	[7067,31263]
Observed	???	???	???	???	???

Table 1: Predictions of new infections by the optimal model and a range given by the space of models equally consistent with the data. March 15 and 16 data arrived at the time of writing. The learned model predicts 657 and 866 new infections on March 15 & 16. The true outcome is toward the lower end of the range, and close to the learned model.

uncertainty, it's a fact of life. There are many models that are equally consistent with the data and together, they give a range of predictions. The model says that if the pandemic is left to its business unchecked, we should expect a total of about 29 thousand new confirmed infections by March 24, and the range is from 28 to 125 thousand.

We can learn a lot from the early dynamics of COVID-19. It's infectious force, virulence, incubation period, unseen infections and predictions of new confirmed cases. All this, albeit within error tolerances, from a simple model and little data. And, the information is useful for planning and health-system preparedness. A side benefit of the predictions in Table 1, if the model is to be trusted, is as a benchmark against which to evaluate public health interventions. If moving forward, observed new infections are low compared to the data in Table 1, it means the interventions are working by most likely reducing β. The US instuted broad and aggressive social distancing protocols starting on or before March 10. If our "lag" $k = 10$ is accurate, and the social distancing protocols have been effective at reducing β, the observed new infections should significantly lag the predictions starting on or before March 20. If we did not have such a target to compare with, it would be hard to evaluate intervention protocols.

Our approach is simple, easy to use, general and works with coarse, aggregated data. We end with some limitations. First, the independent evolution of infection sites only applies to early dynamics. Hence, when the model infections increase beyond some point, and the pandemic begins to saturate the population, a more sophisticated network model that captures the dependence between infection sites would be needed Balcan *et al.* (2009); Hill *et al.* (2010); Salathé and Jones (2010); Keeling and Eames (2005); Chinazzi *et al.* (2020). While we did present an optimal model, it should be taken with a grain of salt because many models are nearly equivalent, resulting in prediction uncertainty. The model and the interpretation of its parameters will change once public health protocols kick in. The model may have to be re-calibrated (for example if β decreases) and the parameters may have to be reinterpreted (for example γ is a virulence only in the self-reporting setting, not for random testing). Lastly, the model was learned on USA data. The learned model parameters may not be appropriate for another society. The virulence could be globally invariant, but it could also depend on genetic and demographic factors like age. One certainly expects the infectious force to depend on the underlying social network and interaction patterns between people, which can vary drastically from one society to another and depending on interventions. Hence, one should calibrate the model to country specific data to gain more useful insights.

References

Abu-Mostafa, Y., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning From Data: A Short Course*. amlbook.com.

Anderson, R. and May, R. (1992). *Infectious diseases of humans*. Oxford University Press.

Bailey, N. (1957). *The mathematical theory of epidemics*. Griffin.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, **106**(51), 21484–21489.

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., y Piontti, A. P., Mu, K., Rossi, L., Sun, K., Viboud, C., Xiong, X., Yu, H., Halloran, M. E., Jr., I. M. L., and Vespignani, A. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, page DOI: 10.1126/science.aba9757.

Cohen, E. (2020). Infected people without symptoms might be driving the spread of coronavirus more than we realized. https://www.cnn.com. (Elizabeth Cohen is Senior Medical Correspondent for CNN).

Hill, A. L., Rand, D. G., Nowak, M. A., and Christakis, N. A. (2010). Infectious disease modeling of social contagion in networks. *PLOS Computational Biology*, **6**(11).

Keeling, M. J. and Eames, K. T. (2005). Networks and epidemic models. *Journal of the Royal Society Interface*.

Kermack, W. and McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A*, pages 700–721.

Kissler, S. M., Tedijanto, C., Goldstein, E. M., Grad, Y. H., and Lipsitch, M. (2020). Projecting the transmission dynamics of sars-cov-2 through the post-pandemic period. medRxiv preprint, Harvard.

Salathé, M. and Jones, J. H. (2010). Dynamics and control of diseases in networks with community structure. *PLOS Computational Biology*, **6**(4).

Wilson, C. (2020). Exclusive: Here's how fast the coronavirus could infect over 1 million americans. TIME web-applet available for scenario analysis.