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 22, 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, including predictions for the number of new infections over time. 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 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 Medical School and adviser to the CDC (see the CNN article, Cohen (2020)). At the time, there was clearly no unimodal opinion. What does the data say? There is a clear need for generating actionable intelligence on the nature of a pandemic from *early* and minimal data.

One approach is scenario analysis. Recently, 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 based 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 level. Such complex scenario analysis requires user input such as infection sites and properties of the contagion. However, a majority of infection sites may be hidden, especially if asymptomatic transmission is significant, and the contagion parameters are *unknown* and must be deduced, perhaps using domain expertise.

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

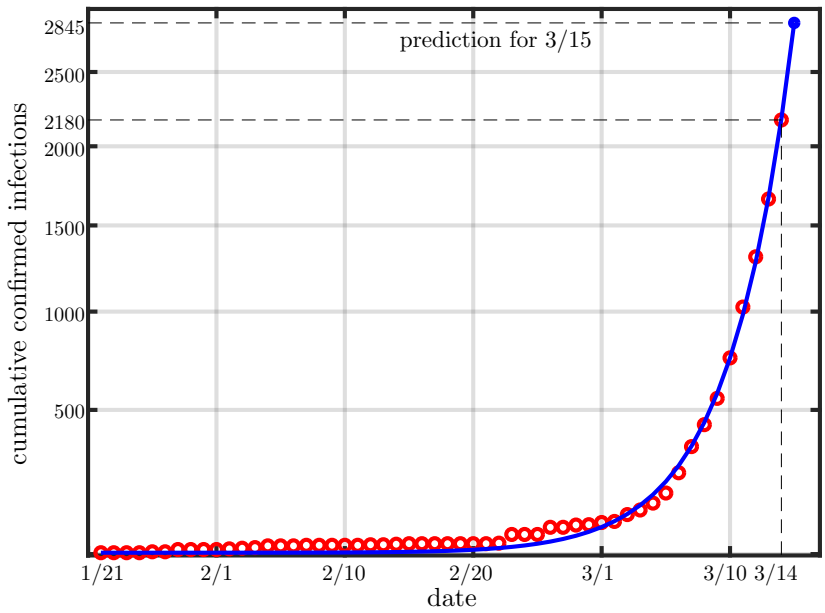

Figure 1: Early dynamics of (USA) COVID-19 and model prediction.

or human input and generates fast actionable insights. Is asymptomatic 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" implies not much information, so quantitative models must be simple and robust to be identifiable 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 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 daily confirmed COVID-19 from January 21 to March 14. Our model is *time stamped*, hence we can provably test it moving forward. 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). Not all serious cases are fatal. Such early intelligence has public health significance: stay calm, the virulence is not astonishingly high, but the spread rate is high, hence one must curtail that spread, or else a health-system can get overloaded.

2 Model, Data and Analysis

Our goal is a simple robust model. 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. This is

important. 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. At time t let there be $C(t)$ 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,

$$M(t) = \beta M(t-1) - \gamma M(t-k) - (1-\gamma)M(t-k-r) + \alpha C(t-1).$$

Here, the second term is the loss of mild cases that transitioned to serious, the third term is the remaining cases that don't transition to serious recovering at some later time r and the fourth term accounts for new infections from confirmed cases. We will further simplify and assume that confirmed cases are fully quarantined and so $\alpha = 0$ and recovery to a non-infectious state occurs late enough to not factor into early dynamics. Our simplified model is:

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

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 assess the quality of the 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 used a combination of root-mean-squared-error and root-mean-squared-percentage-error between observed dynamics and 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.

Let us walk through the process of obtaining the optimal parameters $(\beta^*, \gamma^*, k^*, M_0^*)$. First we determined the incubation period k and the number of initial unseen mild infections M_0. We exhaustively searched over (k, M_0) and for each pair of values, we used gradient descent with 4 million iterations to identify the optimal β and γ for that specific k, M_0. We show the quality of fit for various (k, M_0) in Figure 2(a). The deep-blue region contains 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

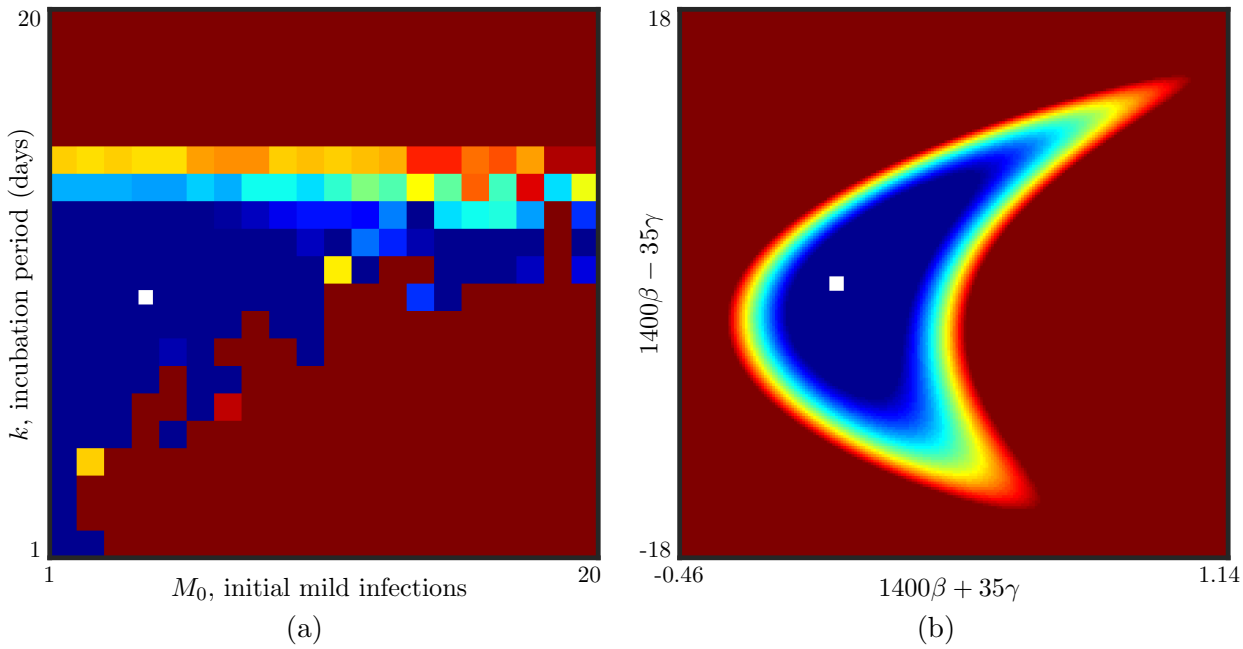

Figure 2: Uncertainty in model estimation. (a) Optimal fit-error over for different choices of incubation period and initial infections. Blue is better fit, red is worse fit. The deep-blue region corresponds to comparable models, within 0.5% of the optimal fit. The white square is the model chosen, $(k^*, M_0^*) = (10, 4)$ 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.

with the data. In making predictions, we should consider all equivalent models to get a range for the predictions that are 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).

2.1 Cross-Sectional Study Across Countries

In the supplementary material we give a cross-sectional study across countries. The different countries have different cultures, social networks, mobility, demographics, as well as different times at which the first infection was reported (the "delay"). The infectious force β has significant variability, and we study how β statistically depends on a number of factors. We find:

- The larger the delay, the larger is β. The more that has been witnessed, the faster the spread. That seems unusual, but is strongly indicated by the data.
- Population density at the infection site has a strong positive effect but the country's population density does not.
- There is faster spread in countries with more people under the poverty level defined as the percentage of people living on less than \$5.50 per day.
- Median age has a strong effect. Spread is faster in younger populations. The youth are more mobile and perhaps also more carefree.
- Wealth and per-capita income have a slight negative effect. Spread is slower in richer countries, perhaps due to some combination of more risk-aversion, higher levels of education and less

March	15	16	17	18	19
New infections	665	866	1130	1475	1924
Range	[657,2861]	[856,3732]	[1114,4867]	[1450,6348]	[1888,8280]
Observed	**777**	**823**	**887**	**1766**	**2988**

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	**4835**	**5374**	**7123**	**???**	**???**

Table 1: Predictions of new infections by the learned 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 and March 17 onward arrived after the time of writing. Blue means in the range and red means outside the range. Note that the new infections (a change in cumulative infections) is less stable than the cumulative infection predictions, which are shown in Figure 3.

reliance on public transportation. Whatever the cause, it does have an impact, but relatively smaller than the other effects.

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$]$

Asymptomatic infectionis strong, around 30%, however it is not that virulent, converting at most 1.2% of mild infections to serious. There is significant uncertainty in the lag, from 1 up to 13 days, and there could have been up to 12 unknown initial infections when the first case was confirmed. As of March 14, the model suggests on the order of a million unknown infections.

We now come to predictions beyond March 14, see Table 1 and Figure 3. Data beyond March 14 was not used to calibrate the model. You may complain about the large uncertainty-range for each prediction. Unfortunately, that is what the data tells us. There is uncertainty, it's a fact of life. There are many models equally consistent with the data, resulting in 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 12. If our "lag" $k = 10$ is accurate, and the social distancing protocols have been effective at reducing β, the observed new infections should drop below the predictions starting on or before March 22. Without such a target to compare with, it would be hard to evaluate intervention protocols.

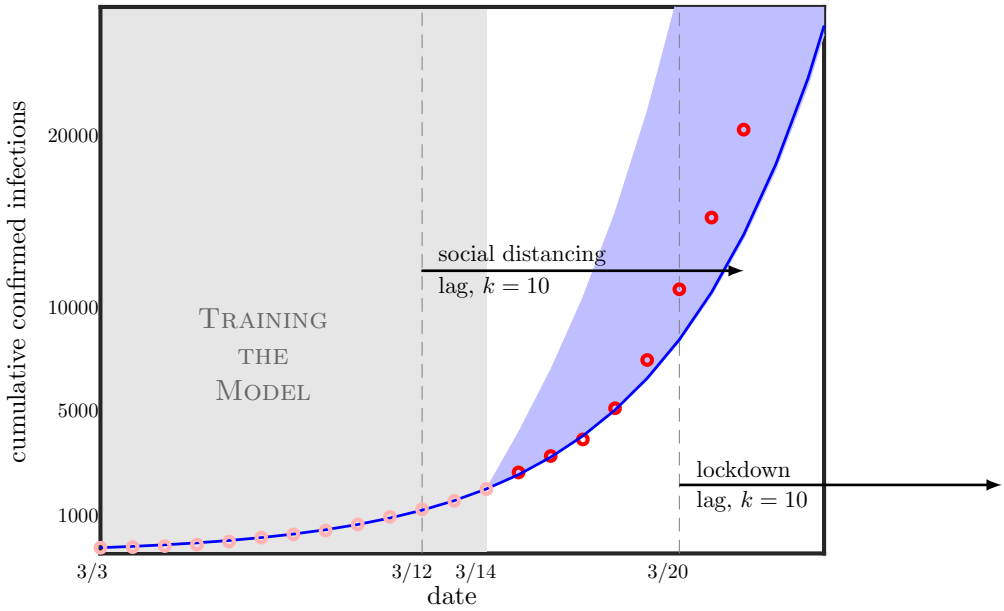

Figure 3: Testing model predictions (the blue envelope indicates the predicted range). The gray region is data used to train the model. A falloff of observed infections away from predicted infections indicates that social distancing is working, and we predict a lag for that of 10 days. (to be seen).

Our approach is simple, 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). It is also possible to build a more general model with an early phase β_1 and a latter phase β_2 (after social distancing). 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, as well as what "serious" means for the culture - that is when do you get yourself checked. In a high-strung society, you expect high virulence since the threshold for serious is lower. 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.

Discussion. Early dynamics allows us to learn useful properties of the pandemic. Later dynamics may be contaminated by human intervention, which renders the data not as useful without a more general model. Are our parameters correct? We are in a unique position to *test* our predictions because our model was *time-stamped* as version 1 of the preprint Magdon-Ismail (2020). The model has provably not changed, and we just added in new test data as it arrives (Figure 3). The model predictions in Figure 3 are in no way forward looking, data snooped or overfitted. Training used

only data before 03/14. The model is holding up.

The Lag, k, and Public Policy. The lag k is important for public policy. Our lag, $k = 10$, says the confirmed infections should fall away from the predictions on about 03/22, because agressive social distancing began on around 03/12. The lag is important due to how public policy can be driven by human psychology as opposed to quantitative models. March 20 saw a 40% increase in new infections compared to the previous day. This led to a significant heightening of vigilance and calls for "shelter in place" protocols. But, by March 20, we still do not know the effectiveness of social distancing protocols from March 12th have been. Hence, such calls to vigilance can be premature. More importantly, suppose strict shelter in place protocols are enacted on March 21 and infections fall the next day. The Human tendency is to associate the improvement to the March 21 strict protocols. But, if the model is correct, the March 21 protocols have nothing to do with the improvement on March 22. Such lags are present in traditional machine learning, for example the delayed reward in reinforcement learning settings. Credit assignment to prior actions in the face of delayed reward is a notoriously difficult task, and this remains so with humans in the loop. Knowledge of the lag helps to assign credit appropriately to prior actions, and the public health setting is no exception.

Acknowledgments. We thank Abhijit Patel, Sibel Adali and Zainab Magdon-Ismail for provoking discussions on this topic.

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).

ECDC (2020). https://www.ecdc.europa.eu. (European Center for Disease Control).

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.

Magdon-Ismail, M. (2020). Machine learning the phenomenology of covid-19 from early infection dynamics. arXiv preprint, https://arxiv.org/abs/2003.07602.

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

Wikipedia (2020). www.wikipedia.com.

Wikipedia, Source: World Bank (2020). List of countries by percentage of population living in poverty. www.wikipedia.com.

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

World-Bank (2017). Adjusted net national income per capita (current us$). https://data.worldbank.org.

World-Data-Info (2015). Average income around the world. https://data.worlddata.info.

A Supplementary Material: Cross-Sectional Country Study

We perform our analysis on early dynamics data available from the ECDC giving infection numbers starting from December 31, 2019 ECDC (2020). We use data for 20 countries selected qualitatively because they appear to have reasonably efficient testing procedures for self-reported cases. We iniclude China data for completeness, even though China dynamics since December 31 are not early dynamics. We show these countries below, together with some demographic data which might determine spread dynamics: Start City; Start and End dates; Delay to first infection in days; Population Density; Median Age (Wikipedia, 2020); Wealth as defined by adjusted net national income per capita (World-Bank, 2017); Average Income (World-Data-Info, 2015); Poverty Level (Wikipedia, Source: World Bank, 2020) defined by living on less than $5.50 per day; Population Density around the first infection site for the country.

A.1 Fitting The Model

Recall the model,

$$\begin{array}{rcll} S(t) & = & S(t-1)+\gamma M(t-k) & S(t)=S(1) \text{ for } 1\le t<k \\ M(t) & = & \beta M(t-1)-\gamma M(t-k) & M(1)=M_0. \end{array} \tag{3}$$

For fixed k, M_0, we must perform a gradient descent to optimize β, γ. Unfortunately, the dependence on β is exponential and hence very sensitive. So if the starting point is not chosen carefully, the optimization gets stuck in a very flat region, and many millions of iterations are needed to converge. Hence it is prudent to choose the starting conditions carefully. To do so, we need to analyze the recursion. First, we observe that the recursion for $M(t)$ is a standalone k-th order recurrence. Hence, we can guess a solution $M(t)=M_0\phi^{t-1}$, for $t\ge 1$, which requires

$$\phi^k-\beta\phi^{k-1}+\gamma=0.$$

We do a perturbation analysis in $\gamma\to 0$. At $\gamma=0$, $\phi=\beta$, so we set $\phi=\beta+\epsilon$, to get

$$(\beta+\epsilon)^k-\beta(\beta+\epsilon)^{k-1}+\gamma=0,$$

which to first order in ϵ is solved by $\epsilon\approx-\gamma/\beta^{k-1}$ and so

$$M(t)\approx M_0\phi^{t-1},$$

where $\phi=\beta(1-\gamma/\beta^k)$. Given this approximation, we can solve for $S(t)$. For $t>k$,

$$S(t)=S(1)+\frac{\gamma M_0(\phi^{t-k}-1)}{\phi-1}.$$

We can independently control two parameters ϕ and γ. We use this to match the observed $S(t)$ at two time points. Since the growth is exponential, we match the end time, $S(T)$ and some time τ in the middle, for example $\tau=\lceil 3T/4\rceil$. Let $\Delta_T=(S(T)-S(1))/M_0$ and $\Delta_\tau=(S(\tau)-S(1))/M_0$. Then,

$$\Delta_T = \frac{\gamma(\phi^{T-k}-1)}{\phi-1}; \tag{4}$$

$$\Delta_\tau = \frac{\gamma(\phi^{\tau-k}-1)}{\phi-1}. \tag{5}$$

Dividing gives $\Delta_T/\Delta_\tau=(\phi^{T-k}-1)/(\phi^{\tau-k}-1)\approx\phi^{T-\tau}$, because $\phi>1$. Let us consider the equation $\kappa=(\phi^r-1)/(\phi^s-1)$, which gives $\phi^r-\kappa\phi^s+\kappa-1=0$, or more generally $\phi^r-\kappa\phi^s+\rho=0$, where $r>s>1$ and $\kappa>\rho\gg 1$. This means $\phi>1$. When $\rho=0$, we have $\phi^{r-s}=\kappa$, so we do a

Country	Code	Start City	Start	End	Delay	Den	Age	Wealth	Income	Pov.	Den-Init
Australia	AU	Melb./Syd.	1/25	3/14	26	3	38.7	41489	53230	1.2	425
Austria	AT	Innsbr./Vien.	2/26	3/14	58	106	44	38748	49310	0.9	3000
Canada	CA	Toronto	1/26	3/14	27	4	42.2	36802	44940	1	4150
China	CN	Wuhan	12/31	2/5	1	145	37.4	6568	9460	27.2	1200
France	FR	Bordeaux	1/25	3/14	26	123	41.4	32672	41080	0.2	5000
Germany	DE	Coppingen	1/28	3/14	29	233	47.1	37791	47090	0.2	972
Iran	IR	Qom/Tehran	2/20	3/14	52	54	30.3	4238	5470	11.6	15500
Ireland	IE	Dublin	3/3	3/14	64	70	36.8	37988	61390	0.7	4588
Italy	IT	Rome	1/31	3/14	32	200	45.5	26537	33730	3.5	2232
Netherl.	NL	Tilburg	2/28	3/14	60	420	42.6	40545	51260	0.5	1852
Norway	NO	Tr. Finn/Oslo	2/27	3/14	59	17	39.2	61865	80610	0.2	1400
Poland	PL	Zielona Gora	3/6	3/14	67	123	40.7	11650	14100	2.1	504
Portugal	PT	Porto	3/3	3/14	64	112	42.2	17188	21990	3	6900
S. Africa	ZA	Gaut./Dur.	3/8	3/14	69	48	27.1	4942	5750	57.1	4000
S. Korea	KR	Daegu	1/20	3/4	21	517	41.8	24028	30600	1.2	2818
Spain	ES	La Gom/Ten.	2/1	3/14	33	93	42.7	23216	29340	2.9	250
Sweden	SE	Jonkoping	2/1	3/14	33	23	41.2	45149	55490	1	2100
Switzer.	CH	Tin./Bas./Zur.	2/26	3/14	58	208	42.4	64307	84410	0	6000
UK	UK	Newcas.Tyme	1/31	3/14	32	274	40.5	34171	41770	0.7	233
USA	US	Seat./Snohom.	1/21	3/14	22	34	38.1	51485	63080	2	3430

Table 2: Comparison of countries used in the study.

perturbation analysis with $\phi^{r-s} = \kappa + \epsilon$, and our perturbation parameter is ϵ. Then, $\phi^r = (\kappa+\epsilon)\phi^s$ and plugging into the equation gives

$$\epsilon = -\frac{\rho}{\phi^s} = -\frac{\rho}{(\kappa+\epsilon)^{s/(r-s)}} \approx -\frac{\rho}{\kappa^{s/(r-s)}}\left(1 - \frac{s}{r-s}\frac{\epsilon}{\kappa}\right).$$

Solving for ϵ gives $\epsilon \approx -\rho(r-s)/((r-s)\kappa^{s/(r-s)} + s)$, which gives

$$\phi \approx \kappa^{1/(r-s)}\left(1 - \frac{(\rho/\kappa)(r-s)}{(r-s)\kappa^{s/(r-s)} - (\rho/\kappa)s}\right)^{1/(r-s)}. \tag{6}$$

For our setting, $r = T-k$, $s = \tau - k$, $\kappa = \Delta_T/\Delta_\tau$ and $\rho = \kappa - 1$. Finally, since ϕ is approximate, we may not be able to satisfy both equations in (5), hence we can instead minimize the mean squared error, which gives

$$\gamma = \frac{(\phi-1)((\phi^r-1)\Delta_T + (\phi^s-1)\Delta_\tau)}{(\phi^r-1)^2 + (\phi^s-1)^2}. \tag{7}$$

We now need to get β which satisfies $\phi = \beta(1-\gamma/\beta^k)$. Again, we do a perturbation analysis, omitting the details, to obtain

$$\beta \approx \phi\left(1 + \frac{\gamma}{\phi^k + (k-1)\gamma}\right). \tag{8}$$

If one wishes, a fixed point iteration starting at the above will quickly approach a solution to $\phi = \beta(1-\gamma/\beta^k)$.

We show the approximate fit on the US data (Figure 4). We show the optimal fit, the initial

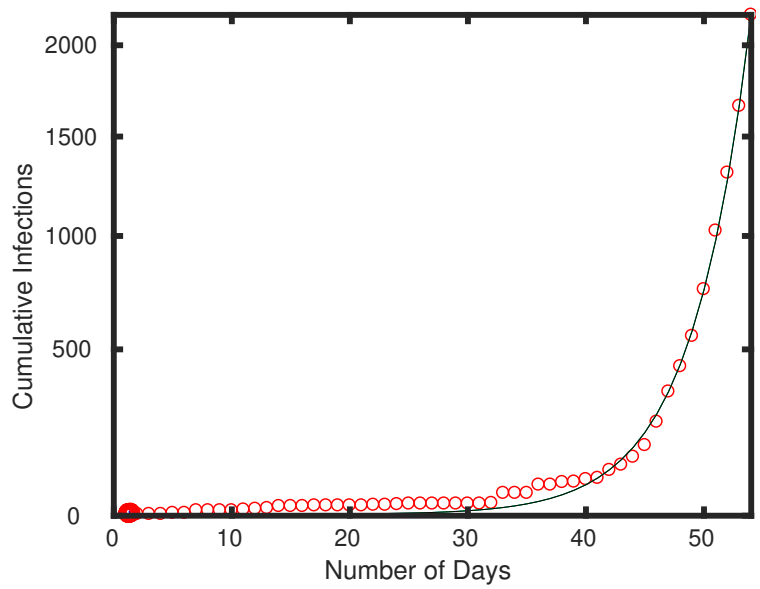

Figure 4: Approximate fit to USA data is essentially on top of optimal fit.

fit using the parameters constructed from (8) and (7). The parameters and fit error are

	β	γ	fit error
optimal	1.306	0.0013282	1.6521
Equations (8) and (7)	1.3055	0.0013423	1.6619

The approximate fit works pretty well, and is certainly good enough to initialize an optimization.

A.2 Country Fits

From the public health perspective, perhaps the most important parameter is β, since actions can be taken to mitigate the spread by reducing β, whereas γ, k and M_0 are somewhat givens for the country. We show the fits in Table 3. As you can see, there is much variability in β.

A.3 Explaining β

We perform a simple statistical analysis to test if β can be explained by any of the country parameters in Table 2. We include the delay as a global explanatory variable, which would account for a global increase in vigilence as time passes and awareness of the pandemic increases. One expects β to decrease with the delay. A table of correlations of β with the various parameters is shown below. For our analysis we use the best case β, although similar results follow from the optimal β.

Dependent var.	Delay	Den	Age	Wealth	Income	Poverty	Den-Init
$\rho(\beta, x)$	0.68	-0.15	-0.51	-0.26	-0.23	0.403	0.52
p-value	0.001	0.52	0.02	0.26	0.32	0.082	0.017

As expected, there is a very significant correlation of β with delay, but in the opposite direction.

- The larger the delay, the larger is β. The more a country has observed, the faster the spread in that country. That seems unusual but seems strongly indicated by the data.

Country	β	[min, max]	γ	[min, max]	k	[min, max]	M_0	[min, max]
Australia	1.143	[1.137,1.18]	0.0047	[0.0024,0.046]	2	[1,3]	10	[1,18]
Austria	1.411	[1.389,2.261]	0.045	[0.022,0.89]	1	[1,1]	20	[1,40]
Canada	1.182	[1.176,1.191]	0.0014	[0.00025,0.0099]	3	[1,3]	10	[1,40]
China	1.312	[1.297,1.332]	5.18	[3.75,8.26]	15	[14,16]	6	[5,9]
France	1.297	[1.293,1.301]	0.05	[0.0128,0.449]	19	[18,20]	7	[1,36]
Germany	1.289	[1.287,1.295]	0.011	[0.0039,0.044]	8	[1,8]	4	[1,10]
Iran	1.596	[1.563,1.709]	0.78	[0.71,0.79]	4	[3,4]	30	[29,40]
Ireland	1.401	[1.372,1.79]	0.083	[0.042,0.62]	2	[2,2]	20	[2,40]
Italy	1.242	[1.237,1.255]	3.88	[3.11,5.37]	19	[18,19]	6	[4,8]
Netherlands	1.387	[1.372,2.28]	0.34	[0.29,1.314]	4	[2,4]	35	[2,40]
Norway	1.461	[1.379,2.935]	0.164	[0.082,1.64]	1	[1,3]	20	[1,40]
Poland	1.476	[1.422,3.542]	0.109	[0.054,2.17]	1	[1,1]	20	[1,40]
Portugal	1.4	[1.356,3.077]	0.0886	[0.044,1.77]	1	[1,1]	20	[1,40]
South Africa	1.563	[1.513,2.65]	0.057	[0.031,1.15]	1	[1,2]	20	[1,37]
South Korea	1.238	[1.231,1.289]	0.98	[0.76,4.65]	20	[14,20]	7	[1,9]
Spain	1.411	[1.405,1.418]	0.127	[0.016,0.63]	20	[20,20]	5	[1,40]
Sweden	1.356	[1.343,1.372]	0.064	[0.0068,0.27]	19	[19,20]	3	[1,40]
Switzerland	1.527	[1.448,2.121]	0.4	[0.267,1.078]	3	[2,3]	12	[2,21]
UK	1.252	[1.248,1.257]	0.68	[0.412,0.878]	19	[17,20]	1	[1,1]
USA	1.306	[1.303,1.309]	0.0014	[0.0006,0.009]	10	[3,12]	4	[1,7]

Table 3: Fit parameters for 20 countries.

- Population density at the infection site has a strong positive effect but the country's population density does not.

- There is faster spread in poorer countries.

- Median age has a strong effect. Spread is faster in younger countries. The youth are more mobile and perhaps also more carefree.

- There is a slight negative effect from wealth and per-capita income. Spread is slower in richer countries. Perhaps this is due to more risk-aversion, perhaps higher levels of education, perhaps less use of public transportation. Whatever the cause, it does have an impact, but relatively smaller than the other effects.

We now use regularized regression to perform a linear model fit to explain β. To make the weight magnitudes meaningful, we normalize the data. We use a leave-one-out cross validation to select the optimal regularization parameter (which happens to be 20). The optimal regularized fit with this regularization parameter gives a new feature

$$X = w_1 \cdot (\text{Delay}) + w_2 \cdot (\text{Pop-Den}) + w_3 \cdot (\text{Age}) + w_4 \cdot (\text{Wealth}) + w_5 \cdot (\text{Income}) + w_6 \cdot (\text{Poverty}) + w_7 \cdot (\text{Pop-Den-Init})$$

The learned weights and the their ranges which yield a cross-validation error within 10% of optimal are shown in the table.

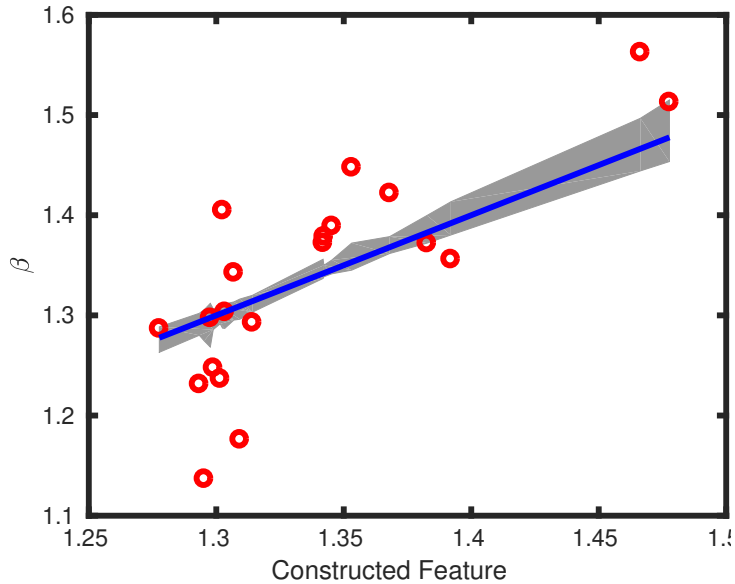

Figure 5: Optimally constructed feature to predict β, within a cross-valudation setting to select the regularization parameter. The gray region is the range of the predicted value for each country. The $R^2 = 0.57$, so the cross-validation based optimal linear feature captures 57% of the residual variance.

Feature	Weight	$[\min, \max]$
Delay	0.030705	[0.023,0.045]
Pop-Den	-0.0014362	[-0.002, 0.00064]
Age	-0.012958	[-0.013, -0.011]
Wealth	-0.004016	[-0.004, -0.0028]
Income	-0.0032445	[-0.0033, -0.0031]
Poverty	0.0109	[0.0089, 0.016]
Pop-Den-Init	0.018305	[0.015, 0.025]

The predictions of β using this feature are also shown in Figure 5. As we observed from the correlations, Delay, Poverty and Population Density at the initial infection site have strong positive weights. Age has a strong negative weight. Wealth and income have weak negative effects, but non-zero. The population density of the country as a whole seems to have no effect, with a weight range that includes 0.