

# Working Paper

## Eliminating COVID-19: A Community-based Analysis

Alexander F. Siegenfeld<sup>1,2\*</sup> and Yaneer Bar-Yam<sup>2</sup>

<sup>1</sup>*Department of Physics, Massachusetts Institute of Technology, Cambridge, MA and*

<sup>2</sup>*New England Complex Systems Institute, Cambridge, MA*

(Dated: March 19, 2020)

We analyze the spread of COVID-19 by considering the transmission of the disease among individuals both within and between communities. COVID-19 can be eliminated if the *community-to-community reproductive number*—i.e. the expected/average number of other communities to which a single infected community will transmit the virus—is reduced to less than one. We find that this community-to-community reproductive number is proportional to the probability that an individual in an infected community will travel to an uninfected community and exponential in the length of the time-delay before community-level action is taken.

Many studies of disease spread consider individuals as the primary unit of analysis, with the reproductive number—i.e. the number of people to whom an infected individual will on average transmit the disease—playing a central role. It is well known that an outbreak can be stopped if interventions reduce this reproductive number to less than 1.

Here, we consider the spread of disease among communities, mediated by its spread among individuals. Central to our analysis is the analogous *community-to-community reproductive number*  $R_*$  [1, 2], i.e. the expected/average number of other communities to which a single infected community will transmit the infection. Our analysis makes no assumptions regarding the size of a community; a community could be as large as a country or as small as a city, so long as disease transmission within communities is far more frequent than disease transmission between communities.

### I. MODEL

The disease is modeled as being transmitted among individuals within a community, with travel allowing the disease to spread between communities. We define a community as *infected* if someone with the infection enters the community. Let  $i_0^c$  be a stochastic factor that roughly corresponds to the initial foothold that the virus gains in community  $c$  conditioning on an infected individual entering the community, with  $i_0^c = 0$  corresponding to the case in which no one was infected or a few people were infected but the outbreak was contained (perhaps through contact tracing and quarantine). If  $i_0^c = 0$ , the outbreak spreads no further within or between communities, and so we can write  $i_c(t) = 0$ , where  $i_c(t)$  is the number of active infections in community  $c$  as a function of time.

If the infection is not contained (i.e.  $i_0^c \neq 0$ ), the number of active infections is modeled as growing with time  $t$  at an exponential rate  $e^{r_c t}$ . After some time  $T_c$  (the delay in response), the community implements aggressive

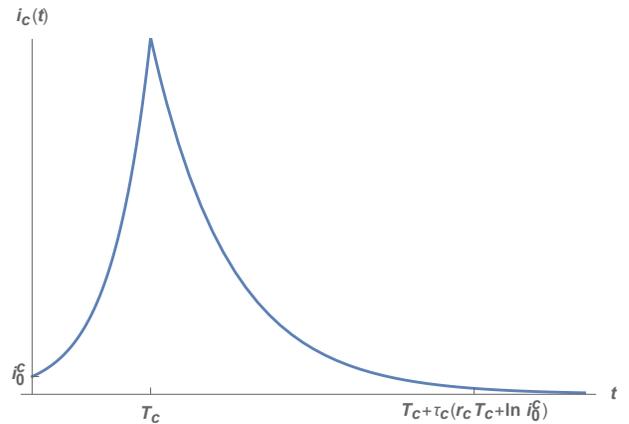


FIG. 1. The number of active infections in a community as a function of time. After time  $T_c$ , aggressive social distancing measures are implemented. They must remain in place for a duration greater than  $\tau_c(r_c T_c + \ln i_0^c)$  (see eq. (2)). Thus, the longer the community waits to enact the measures, the longer the total amount of time they must remain in place.

social distancing measures that cause the number of active infections to decay as  $e^{-t/\tau_c}$ , where  $\tau_c$  is the amount of time that it takes for the number of active infections to drop by a factor of  $e$ .<sup>1</sup>

The number of active infections  $i_c(t)$  (see fig. 1) can therefore be written as:

$$i_c(t) = \begin{cases} i_0^c e^{r_c t} & t \leq T_c \\ i_0^c e^{r_c T_c} e^{-(t-T_c)/\tau_c} & t \geq T_c \end{cases} \quad (1)$$

<sup>1</sup> The social distancing measures are described as aggressive, since they must be comprehensive enough to cause the reproductive number to drop below 1. The basic reproductive number  $R_0$  for COVID-19 is estimated at approximately 2.5 (see section II), and the spread of the virus must be cut by a factor greater than  $R_0$ . Simulations from the Institute for Disease Modeling show that a 75% contact reduction will result in an exponential decrease in the number of active infections after a time-delay due to the incubation periods and durations of existing infections [3].

\* Corresponding author. Email: asiegenf@mit.edu

The social distancing measures can be lifted once there are no remaining active infections in the community or once all active infections have been contained. Solving for  $i_c(t) = 1$  (assuming  $i_0^c \neq 0$ ) yields a duration of the aggressive social distancing measures of

$$\tau_c(r_c T_c + \ln i_0^c) \quad (2)$$

plus some additional time at the end to ensure that there are no remaining hidden cases. As the number of cases becomes increasingly small, contact tracing may become increasingly effective and hasten the drop of  $i_c(t)$  to 0.

Each infected community  $c$  infects a currently uninfected community with a probability rate proportional to the number of active infections  $i_c(t)$  times the probability rate  $p_c$  that an infected individual will travel to an uninfected community. The number of new infected communities spawned by community  $c$  can thus be modeled as a Poisson process with rate  $i_c(t)p_c$ . This modeling assumption overestimates the spread of the disease to new communities by counting a single new community that has been infected multiple times as multiple new infected communities. Note that a single new community that is infected by more than one other community is also counted as multiple new infected communities. If most communities are uninfected and the virus is being contained, an infected community will infect on average less than one other community, and thus there is a negligible probability that the same one will be infected twice. However, if many communities are infected and/or the number of infected communities is exponentially growing, this analysis serves only as an upper bound, as it neglects saturation effects. In order for the disease to be eliminated, the former scenario must be attained, and thus the Poisson assumption does not affect our conclusions.

Let  $p_0^c$  be the per capita probability rate before time  $T_c$  of individuals in community  $c$  traveling to other communities and  $p_1^c$  be the probability rate afterwards ( $p_1^c$  will be less than  $p_0^c$  if travel is discouraged and/or restricted at the time at which social distancing measures are implemented). The number of new communities that are infected by community  $c$  will then be a Poisson random variable with a mean of

$$i_0^c p_0^c \int_0^{T_c} e^{r_c t} dt + i_0^c p_1^c e^{r_c T_c} \int_0^{\infty} e^{-t/\tau_c} dt \quad (3)$$

$$= i_0^c \left( p_0^c \frac{e^{r_c T_c} - 1}{r_c} + p_1^c \tau_c e^{r_c T_c} \right) \quad (4)$$

Taking the expected value over  $i_0^c$  yields

$$R_*^c = \mathbb{E}[i_0^c] \left( p_0^c \frac{e^{r_c T_c} - 1}{r_c} + p_1^c \tau_c e^{r_c T_c} \right) \quad (5)$$

where  $R_*^c$  is the community-to-community reproductive number for community  $c$ , i.e. the expectation of the number of communities that community  $c$  will infect if it is infected. As the parameters  $r_c$ ,  $\tau_c$ ,  $T_c$ ,  $\mathbb{E}[i_0^c]$ ,  $p_0^c$ , and  $p_1^c$

may differ from community to community, so will  $R_*^c$ . If the interventions are fast enough and strong enough such that  $R_*$ , the average value of  $R_*^c$  over a set of communities with each community weighted by its probability of being infected, is less than 1, then the outbreak will not be self-sustaining within that set of communities.

A set of communities can thus exist in one of two phases or regimes: a regime in which over time the number of infected communities exponentially decreases to zero and a regime in which over time the number of infected communities exponentially increases until it saturates as an endemic disease or reaches burnout. Thus, small deviations from the assumptions that led to the formula for  $R_*^c$  will not matter so long as they do not change which regime the system is in, i.e. whether or not  $R_* < 1$ . It is this universality [4] that allows us to understand whether or not any disease will spread, even if it is not possible to precisely describe the details of disease transmission and social connectivity.

## II. PARAMETER ESTIMATION

In order to better understand the extent of the measures required to achieve  $R_* < 1$ , we estimate the values of the parameters in eq. (5).

The doubling time can vary from location to location, but using the number of confirmed cases in China by date of symptom onset (rather than by date of diagnosis) [5] gives a doubling time of 3.04 days in the period leading up to the Jan. 23 lockdown, which corresponds to  $r = 0.228 \text{ day}^{-1}$  (see fig. 2).<sup>2</sup>

From this growth rate before the Jan. 23 lockdown, the basic reproductive number  $R_0$  can be calculated from the distribution of generation intervals [10].<sup>3</sup> Empirical data from various sources support a distribution of generation intervals with a mean of approximately 4.0 to 4.7 days [11–16], which yield upper bounds of  $R_0 < e^{4.0r} = 2.5$  to  $R_0 < e^{4.7r} = 2.9$ . These upper bounds assume all transmission occurs at the mean generation interval; the spread of generation intervals given

<sup>2</sup> Some studies estimate the doubling time for COVID-19 at approximately 7 days [6, 7], but even a 5 day doubling period is implausibly long, given that in various countries, even with some preventative measures, the number of infections has increased by far more than a factor of 64 over 30 days [8]. Part of the difficulty in estimating the doubling time from the initial period of transmission is that ‘super-spreader’ events may play an important role in the transmission process. The presence of super-spreader events indicates that the transmission process may be fat-tailed and therefore standard statistical approaches may underestimate the rate of spread when the total number of cases is still small [9].

<sup>3</sup> Empirically, we generally observe the distribution of serial intervals (the times between the onsets of symptoms in two successive cases in a transmission chain) rather than the distribution of generation intervals (the times between two successive infections in a transmission chain). The means of the two distributions should, however, be the same.

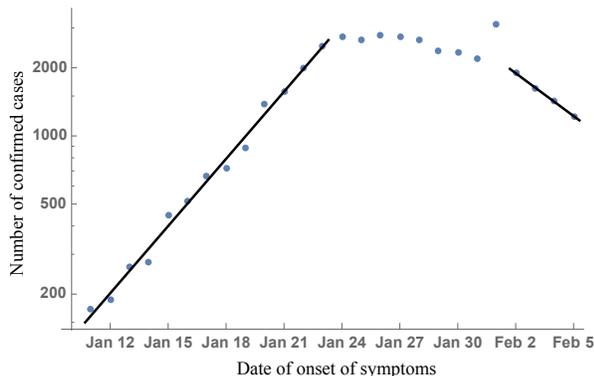


FIG. 2. Log plot of the daily number of confirmed cases in China by date that these patients self-reported as the onset of their symptoms. The best OLS linear fits to the natural log of the number of cases are shown: For Jan. 11-23 (up until the lockdown), the slope is 0.228 ( $R^2 = 0.991$ , 95% confidence interval [0.214, 0.242]), which corresponds to a doubling time of 3.04 days. For Feb. 2-5 the slope is  $-0.145$  ( $R^2 = 0.999$ , 95% confidence interval  $[-0.160, -0.131]$ ), which corresponds to a halving time of 4.78 days. An exponential decline is not seen until Feb. 2, likely because cases with onsets of symptoms between Jan. 24 and Feb. 1 include infections that occurred both before and after the lockdown due to the wide range of incubation periods. Data are from [5], which includes cases diagnosed through Feb. 11. Not pictured: There is a drop in cases with onsets of symptoms after Feb. 5, likely due to many of those cases being diagnosed after Feb. 11. It should be noted that the number of cases by date of diagnosis (not pictured) continues to increase through Feb. 4, indicating that in general the date of diagnosis substantially lags the date of onset of symptoms, which itself lags the date of infection. Thus, a considerable amount of time can pass before even an immediate decrease in the infection rate can be observed.

a fixed mean results in lower values for  $R_0$  given a fixed  $r$  (or higher values for  $r$  given a fixed  $R_0$ ). For instance, the serial interval data with a mean of 4.7 days best fit a lognormal distribution (S.D.=2.9 days) [11], which yields  $R_0 = 2.5$ .<sup>4</sup> This short mean generation time of 4.0 to 4.7 days may explain the difficulty of containing the virus through contact tracing alone.

The values of  $\tau_c$  that can be achieved depend on the effectiveness of the social distancing measures. The data from China (see fig. 2) indicate a halving time of as little

<sup>4</sup> This value was obtained by approximating the generation interval distribution by the serial interval distribution. It is consistent with some previously reported  $R_0$  values that were based on a mean generation interval overestimated as similar to that of SARS (8.4 days)/MERS (7.6 days) [17] because the doubling time was also overestimated by a similar factor.

as 4.78 days is achievable, which corresponds to  $\tau = 6.9$  days.

$\mathbb{E}[i_0^c]$  is the expected “effective” number of people an infected traveler will infect while visiting community  $c$ , taking into account containment efforts.<sup>5</sup> We estimate  $\mathbb{E}[i_0^c] = R_0$ ; the degree to which  $\mathbb{E}[i_0^c]$  differs from  $R_0$  depends on how likely a typical traveler is to transmit the virus relative to a typical resident, as well as on the effectiveness of contact tracing and other containment efforts.

The value of  $p_0^c$  depends on the frequency of travel out of community  $c$ . There is some choice in how to model the partition of a population into a set of communities. In general, the larger the communities, the lower the frequency of per capita travel out of them but the more homogeneous the application of the aggressive social distancing measures. Considering a set of communities within the U.S. that are large enough such that travel between the communities is predominantly by flight yields a per capita travel rate of 0.004 flights out of a community per person per day.<sup>6</sup>

The values above allow us to use eq. (5) to determine  $R_*$  as a function of the time-delay before aggressive social distancing measures are enacted, as shown in fig. 3. Note that the time-delay is measured from the time at which exponential growth begins to occur—which could be as early as the first infection transmitted within the community if containment is not successful—not the time at which exponential growth is first measured.

### III. THE IMPACT OF INTERVENTIONS

Regardless of the values of the parameters, i.e. regardless of whatever the current value of  $R_*$  is,  $R_*$  can be reduced by a number of interventions:

- A reduction in travel from community  $c$  results in a proportional (linear) reduction in  $R_*^c$  through  $p_0^c$  and  $p_1^c$ .
- Improvements in testing, contact tracing, and quarantine result in a proportional reduction in  $R_*^c$  through  $\mathbb{E}[i_0^c]$ . Such improvements also reduce  $r_c$  and  $\tau_c$ .
- The preemptive reduction of large events such as conferences (as well as general social distancing)

<sup>5</sup> For instance, if the outbreak is contained such that exponential growth never occurs, the effective number of people infected by the traveler is zero, even if the traveler did infect some community members.

<sup>6</sup> This estimate is obtained by dividing the 1.01 billion total passengers traveling by plane to, from, or within the U.S. in 2018 [18] by the 2018 U.S. population and the number of days in 2018, and then also dividing by 2 so that only flights out of and not into a community are counted. Using this estimate for  $p_0^c$  assumes that the probability that an infected individual will travel equals that of the general public.

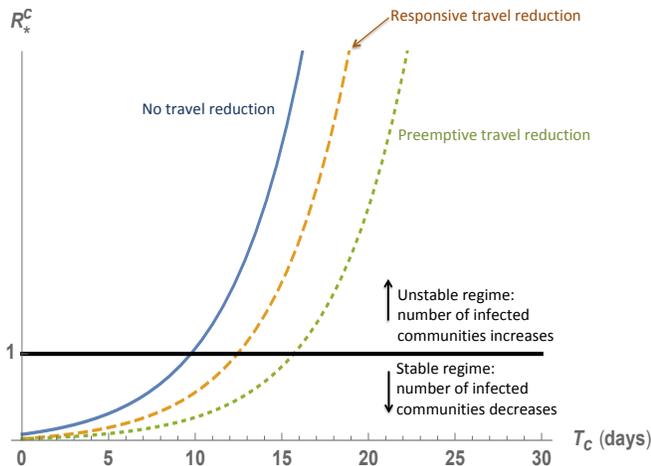


FIG. 3. Dependence of  $R_*^c$  (the average number of communities to which community  $c$  will transmit the disease) on  $T_c$  (the time-delay before the social distancing measures are enacted). If  $R_*$  (a weighted average of  $R_*^c$ ) is less than 1, the number of infected communities will exponentially decrease and the disease will be eliminated (and the smaller  $R_*$  is, the faster it will be eliminated); otherwise, the number of infected communities will increase until saturation.

**Parameter values:** All curves use  $\mathbb{E}[i_0^c] = 2.5$ ,  $\tau_c = 6.9$  days, and  $r_c = .228 \text{ day}^{-1}$ . Solid curve: no travel reduction,  $p_0^c = p_1^c = 0.004 \text{ day}^{-1}$ . Dashed curve: 4-fold (responsive) travel reduction after time  $T_c$ ,  $p_0^c = 0.004 \text{ day}^{-1}$  and  $p_1^c = 0.001 \text{ day}^{-1}$ . Dotted curve: general (preemptive) 4-fold travel reduction:  $p_0^c = p_1^c = 0.001 \text{ day}^{-1}$ .

reduces the probability of a super-spreader event as well as general transmission, reducing both  $\mathbb{E}[i_0^c]$  and  $r_c$ .

- Reductions in  $r_c T_c$  not only exponentially reduce  $R_*^c$  but also linearly reduce the amount of time for which social distancing measures must remain in place.
- Augmenting social distancing measures (after time  $T_c$ ) decreases  $\tau_c$ , which results in a proportional decrease in both  $R_*^c$  and the time for which the distancing measures must remain in place.

We conclude with a few comments. First, without the timely implementation of aggressive social distancing measures, restricting travel from infected communities serves only to delay the spread of the outbreak. But when a reduction in travel is coupled with the social distancing measures, the travel reduction will not only delay the spread of the outbreak but will also in some cases be the deciding factor in determining whether or not the outbreak is eliminated. And even if  $R_* < 1$  can be achieved

without reducing travel, travel reductions can, by further reducing  $R_*$ , greatly decrease the duration of the outbreak (and therefore also decrease the total number of illnesses and deaths).

Second, the most important and easiest parameter to control is  $T_c$ : the amount of time delay between the time the infection takes hold in the community and the time that aggressive social distancing measures are implemented. Since  $R_*^c$  grows exponentially with  $T_c$ , it is of paramount importance that communities and governments act as soon as possible. Once the infection has taken hold in a community, exponential growth ensures that it is only a matter of time before the infection becomes widespread. Thus government action is inevitable, and delaying action not only linearly increases the expected total amount of time for which the distancing measures will need to remain in place but also exponentially increases the probability that another community will become infected.

Third, because  $R_*^c$  depends exponentially on  $T_c$ , each additional increase in  $T_c$  becomes increasingly costly. In other words, the longer a community has already waited to take aggressive social distancing measures, the more important it becomes to avoid further delay. It is important to note, however, that there is no advantage to delaying at all: immediately implementing aggressive social distancing measures as soon as there is evidence that the disease is actively spreading within the community will reduce the total amount of time for which such measures will need to be implemented, exponentially reduce the number of infections within the community, and exponentially reduce the probability of infecting another community. While the risk to individuals may be low in the early stages of a community outbreak, the collective risk at the community and global levels is extremely high.

Finally, just as transmission within a community is an exponential process in which the need for immediate action is not always apparent (due to a deceptively small number of initial cases), the transmission between communities is also such an exponential process. At first the number of infected communities is deceptively small, but without rapid action this number exponentially grows. The sooner a set of communities decides to adopt a protocol that reduces  $R_*$  below 1, the shorter the amount of time between the adoption of the protocol and the elimination of the disease.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1122374 and the Hertz Foundation. A.F.S. thanks Maxim Rabinovich for helpful early discussions about the model.

- 
- [1] Ball, F., Mollison, D. & Scalia-Tomba, G. Epidemics with two levels of mixing. *The Annals of Applied Probability* 46–89 (1997).
- [2] Colizza, V. & Vespignani, A. Invasion threshold in heterogeneous metapopulation networks. *Phys. Rev. Lett.* **99**, 148701 (2007).
- [3] Klein, D., Hagedorn, B., Cliff Kerr, H. H., Bedford, T. & Famulare, M. Working paper - model-based estimates of COVID-19 burden in King and Snohomish counties through April 7, 2020. *Institute for Disease Modeling* (2020).
- [4] Siegenfeld, A. F. & Bar-Yam, Y. An introduction to complex systems science and its applications (2019). arXiv:1912.05088.
- [5] Wu, Z. & McGoogan, J. M. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72314 cases from the Chinese Center for Disease Control and Prevention. *JAMA* (2020).
- [6] Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* **395**, 689–697 (2020).
- [7] Li, Q. *et al.* Early transmission dynamics in Wuhan, China of novel coronavirus-infected pneumonia. *New England Journal of Medicine* (2020).
- [8] Worldometer. COVID-19 coronavirus outbreak. <https://www.worldometers.info/coronavirus/>. Accessed 2020-3-17.
- [9] Taleb, N. N. How much data do you need? An operational, pre-asymptotic metric for fat-tailedness. *International Journal of Forecasting* **35**, 677 – 686 (2019).
- [10] Wallinga, J. & Lipsitch, M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences* **274**, 599–604 (2007).
- [11] Nishiura, H., Linton, N. M. & Akhmetzhanov, A. R. Serial interval of novel coronavirus (COVID-19) infections. *International Journal of Infectious Diseases* (2020).
- [12] Du, Z. *et al.* The serial interval of COVID-19 from publicly reported confirmed cases. *medRxiv* (2020).
- [13] Zhao, S. *et al.* Estimating the serial interval of the novel coronavirus disease (COVID-19): A statistical analysis using the public data in Hong Kong from January 16 to February 15, 2020. *medRxiv* (2020).
- [14] You, C. *et al.* Estimation of the time-varying reproduction number of COVID-19 outbreak in China. *Available at SSRN 3539694* (2020).
- [15] Tindale, L. *et al.* Transmission interval estimates suggest pre-symptomatic spread of COVID-19. *medRxiv* (2020).
- [16] Ganyani, T. *et al.* Estimating the generation interval for COVID-19 based on symptom onset data. *medRxiv* (2020).
- [17] Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklöv, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of travel medicine* (2020).
- [18] U.S. Department of Transportation. 2018 traffic data for U.S. airlines and foreign airlines U.S. flights. <https://www.bts.dot.gov/newsroom/2018-traffic-data-us-airlines-and-foreign-airlines-us-flights>. Accessed 2020-3-12.