Forecasting the Spread of COVID-19 under Different Reopening Strategies[*]

Meng Liu Raphael Thomadsen Song Yao

Olin Business School
Washington University in St. Louis

This draft: May 22, 2020

Abstract

We combine COVID-19 case data with demographic and mobility data to estimate a modified SIR model for the spread of this disease in the United States. We find that the incidence of infectious COVID-19 individuals has a concave effect on contagion, as would be expected if people have inter-related social networks. We also demonstrate that social distancing and population density have very large causal effects on the rate of contagion. The social distancing that occurred in late March and April substantially reduced the number of COVID-19 cases in the U.S. However, the concave contagion pattern means that, when social distancing measures are lifted, the growth rate is considerable but will not be the exponential growth predicted by standard SIR models. Furthermore, counties with the lowest population density could likely return to no social distancing and avoid high levels of contagion. We also find that higher temperatures do reduce the spread of COVID-19, but by much smaller amounts than social distancing and population density. After confirming that our model is able to forecast the rate of contagion well in early May, we forecast rates of new cases for COVID-19 under different social distancing norms. We find that if social distancing is returned to normal levels then there will be a massive increase in the cases of COVID-19 in the U.S., about double of what would be obtained if the US only restored to 50% of the way to normalcy. Most of this increase would be driven by cases in the most-densely populated counties.

Keywords: COVID-19, SIR models, Social distancing

[*] We thank SafeGraph Inc. for providing the mobility data used in our analysis. All authors contribute equally and listed alphabetically. Please contact Liu (mengl@wustl.edu), Thomadsen (thomadsen@wustl.edu), or Yao (songyao@wustl.edu) for correspondence.

1 Introduction

As COVID-19 spreads across the world and the United States, governments and individuals have worked to slow the growth of the disease by reducing the extent to which people leave their homes. In the United States, these actions have largely been acted on by households who have voluntarily stayed home unless they needed to travel, but they have also been bolstered by orders of local and state governments. These orders have occurred over different time periods, and have taken many different forms, but they have had a similar flavor of limiting gathering sizes, closing schools, and shutting down non-essential businesses or shifting their operations to a contact-free experience. That said, many areas of the country never were given any form of stay-at-home orders.

Ultimately, the purpose of the stay-at-home orders is to reduce the amount of contact between people in order to slow the growth of COVID-19, which is thought to be spread primarily through droplets that require being within a relatively small distance of an infected person. In this paper, we first measure the extent to which social distancing causally reduces the speed at which COVID-19 spreads. We then run simulations of how COVID-19 will spread over time under different policy regimes.

We find that COVID-19 spreads less than proportionately with the spread of the disease. We also observe that social distancing during late March and April significantly reduced the spread of the disease. We find that higher temperatures will slow the growth of COVID-19 in the summer, but by much less than the impact of social distancing. Higher population density leads to an increased spread of COVID-19, as well.

Our model gives good out-of-sample forecasts of the disease for the two and half weeks after the end of our mobility data, assuming that the nearly 50% return to normalcy observed at the end of April (as compared to the observed peak social distancing levels) continues. We forecast that completely opening up the country to 100% of the pre-shutdown levels of social interaction will lead to 4 million additional COVID-19 cases (officially diagnosed) by the end of September 2020, corresponding to a doubling of the cases that we would expect were we continuing with a median-level social distancing between the peak and the per-shutdown levels. However, there is a great heterogeneity among counties, and according to our simulations 44% of the counties could open up while still experiencing a low infection rate less than 0.1% over a 3-month period.

The model we estimate is a simplified version of an SIR model. We assume that

$$y_{i,t} = R_{i,t} S_{i,t} \left(Y_{i,t-2} - Y_{i,t-8}\right)^{\omega} \tag{1}$$

where $y_{i,t}$ is the number of individuals who are infected in county i on day t, $R_{i,t}$ is the rate at which infectious individuals in the county transmit the disease, $S_{i,t}$ is the percentage of the county population that is susceptible to COVID-19 (i.e., the share of people who have not yet had COVID-19), and $Y_{i,t}$ is the number of cumulative individuals who have been infected up until day t. The $Y_{i,t-2} - Y_{i,t-8}$ term reflects our assumption that infected individuals are contagious from the second day after they catch the virus through the seventh day, leading to a serial interval of 4.5

days Nishiuram et al. (2020). This treatment of the infectious population is an approximation to the way SIR models are typically modeled, where the infectious population is modeled as a stock that has an outflow at a constant rate. This assumption makes the estimation much easier with the large number of fixed effects we include in our model, and as a practical matter this assumption only has a minimal impact on our estimates for the contagion of COVID-19. In the supplemental appendix we show that we get extremely similar results if we take the time of contagiousness to be 14 days instead of 6 days.

The main difference between this model and a standard SIR model is that a standard SIR model constrains $\omega = 1$. We show in the supplemental appendix that the estimated model with this prediction does not perform well out of sample. We instead find that $\omega < 1$. This shows that the marginal impact of one more sick person diminishes as more and more people are sick. There are several reasons why this may be expected, with the greatest reason being that contagious individuals may end up endangering many of the same group of unexposed individuals. One might expect this to be the case if people often have the same or overlapping groups of friends or aquantinances. We see some of this directly when, for example, cases are clustered within households, nursing homes, or places of work. In the accompanying supplemental appendix we present a networking model and show that we would get $\omega < 1$ if people have interconnected networks of contact.

In order to better understand the variation of the rate of contagion, we allow $R_{i,t}$ to vary according to a number of factors instead of treating it as a constant parameter. Thus, we model

$$R_{i,t} = \exp\left(\alpha_i + \beta_t + \lambda d_{i,t} + \theta h_{i,t} + \varepsilon_{i,t}\right) \tag{2}$$

This specification implies that transmission rates can differ across counties (the county fixed effects α_i reflect different population densities and also different demographic compositions), time periods (date fixed effects β_t are included mostly to accommodate different rates of testing and also the different rates of reporting that happen on weekdays vs weekends), levels of social distancing $d_{i,t}$, and different temperatures, $h_{i,t}$, the impact of which has been debated (Wang et al. 2020a, Oliveiros et al. 2020, Wang et al. 2020b). The social distancing data, $d_{i,t}$, is based on cellphone GPS location data that are provided by SafeGraph, and are available for free to researchers studying COVID-19. We measure social distancing as the fraction of phones that stay exclusively at home during a given day.

The $\varepsilon_{i,t}$ term is our statistical error term. Equation 1 is estimated by taking the logarithm of both sides, with the details in the appendix. Note that the social distancing level by individuals, as well as social distancing regulations, are not determined in a vacuum. Rather, we observe that people social distance more in areas that are harder hit by COVID-19. Thus, the $\varepsilon_{i,t}$ term may be correlated with the social distancing measures, causing a biased underestimation of the impact of social distancing on slowing the spread of the disease. We control for this statistical bias by estimating the model using an Instrumental Variables technique, where we use the amount of rain as an instrument for social distancing. Specifically, we assume that rain directly shifts the level of social distancing, but is not correlated with $\varepsilon_{i,t}$.

	(1) IV	(2) OLS
Dependent Variable	Log(Infected in County i	Log(Infected in County i
	on Date t)	on Date t)
Social Dist. Level in	-5.445***	-0.088
County i on Date t	(1.656)	(0.194)
Infectious Individuals in	0.470***	0.449***
County i on Date t	(0.014)	(0.012)
Avg. Temperature ($°C$) of	-0.009***	-0.003*
County i on Date t	(0.002)	(0.002)
County Fixed Effects	Yes (Mean=0.262)	Yes
Date Fixed Effects	Yes (Mean=3.835)	Yes
Observations	73,713	73,713
R_squared	0.61	0.62
Counties	2704	2704

*** p<0.01, ** p<0.05, * p<0.1

Table 1: **Estimation of a Modified SIR Model.**

To calibrate the model parameters, we use county-level officially confirmed COVID-19 daily case data of 2,704 US counties or county-equivalents from February 3 to April 28. We append the data with the county-daily-level weather data as well as cellphone mobility data provided by SafeGraph. The results are presented in Table 1. Column (1) in the table shows our final result. We find that social distancing indeed decreases the growth rate of COVID-19: Moving from the observed mean pre-COVID level of social distancing (0.25) to the post-COVID peak level (0.40), the magnitude of R is reduced by 56%. Thus, social distancing has a very substantial impact on the spread of the disease. Note that if we had not instrumented for social distancing we would not find an significant social distancing effect on the spread of COVID-19, as would be expected given the statistical bias issue aforementioned. Those biased results appear in the second column of Table 1 for comparison.

We also find that the exponent on the number of contagious people is 0.47, significantly lower than 1, the exponent assumed in a standard SIR model. This shows that there is a strongly concave relationship between the number of infected people and the rate that the disease spreads. This level of concavity also implies that while initial outbreaks of COVID-19 would expand exponentially, they will quickly turn to a slower rate of growth. The growth looks linear or even plateauing while being plotted cumulatively, although the disease will persist for a long period of time and continue building a substantial number of cases. This may explain why the growth rate of the number of COVID-19 cases has recently slowed down considerably after a quick take-off, and yet this growth has persisted. A quick look at the diffusion paths of many countries and US states shows a pattern that resembles such a trend, which is also consistent with our forecasts below. We also find the higher temperatures may slow the spread of the virus, but with a much smaller impact.

Most of the variation of the contagion rates $R_{i,t}$ is captured by our county-level fixed effects, α_i in Equation 2. In order to understand the drivers of the contagion rates, we run a regression of

Dependent Variable	County Fixed Effect
Log(Pop. Density)	0.3930***
(People/Sq. Miles)	(0.0102)
Percentage of African	1.4175***
American Residents	(0.1008)
Percentage of Hispanic	1.3755***
Residents	(0.1051)
Percentage of Commuters	5.4436***
using Pub. Transportation	(0.6639)
Log(Median Income)	1.1244***
(in U.S. dollars)	(0.0626)
Percentage of Senior	1.0473*
Residents ($\geq$70yrs)	(0.5564)
Percentage of Children	0.7803
Residents (<18yrs)	(0.5089)
Constant	-14.1108***
	(0.6923)
R_squared	0.62
Counties	2,703

*** p<0.01, ** p<0.05, * p<0.1

Table 2: **Factors Affecting County Fixed Effects of Rate of Contagion**

the county fixed effects α_i against county-level demographics. Hansen (2007) shows that when the number of fixed effects are large then these coefficients can be treated as data for the purpose of statistical inference. The results are in Table 2. We find that population density is a crucial factor influencing the spread of the disease. In fact, in some areas with low enough population densities COVID-19 would be expected to never flare up beyond a very small base level as we discuss below in our forecasting. We also observe that greater concentrations of Black and Hispanic residents and public transit commuters are associated with higher contagion rates. Interestingly, higher median incomes are related to more contagion. We are uncertain what drives this result: one possibility is that people with higher incomes may be interacting more with nearby cities that have more outbreaks. We also include the fraction of the population that consists of seniors (age 70+) and children (below age 18). Seniors are marginally more likely to spread the disease, but children show no sign of having a lower rate of infecting people, confirming Danis et al. (2020).

Using our model, we simulate future cases beyond our sample. First, to examine how our model performs, we predict the out-of-sample case numbers from the end of our data period up to May 16, 2020, under different social distancing assumptions. We start by forecasting the cumulative COVID-19 cases if each county continued the social distancing at the levels observed at the end of April. Assuming the observed February level as normalcy, the end-of-April level is at the 50% level between the peak lock-down level and normalcy (so we say that such a level is at 50% of normalcy).

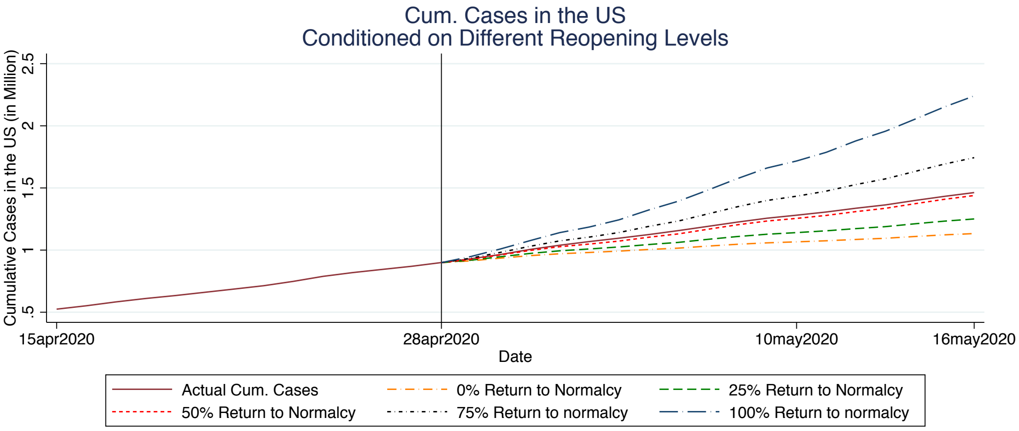

Figure 1: Out-of-Sample Prediction under Different Reopening Strategies. The vertical line indicates the last day of our data sample.

We also implement the same exercise under several other social distancing benchmarks. These benchmarks are defined specifically in the supplemental appendix. The results appear in Figure 1. We observe that our model forecasts the pattern of disease contagion well if the level of social distancing in early May remained at the level seen at the end of April (50% return to normalcy).

Finally, we forecast how the disease will evolve up to September, 2020, under different reopening strategies. The cumulative and daily cases appear in Figures 2 and 3, respectively. The cyclicality that is observed in Figure 3 reflects the variation we observe in the weekly data, which may reflect different reporting delays or different social-distancing behavior due to the day-of-the-week effect. The forecast shows that social distancing matters, but that the impact of increased mobility becomes higher as we move closer to normalcy. If social distancing is eliminated, we observe that the largest effects will be felt in the first two months. This occurs because of the shrinkage of the uninfected population in each county. Note that cases will be elevated at almost double the daily rate that we would observe under a 50%-return-to-normalcy even into the September of 2020, when cases are likely to reach an almost steady weekly level. Our estimates suggest that it will be difficult to return to school and normalcy in the fall of 2020 without sparking a large outburst of COVID-19. Ultimately, moving from the 50%-return-to-normalcy we observe at the end of April to a full return to normalcy will lead to 4 million additional confirmed cases of COVID-19. If we assume that the confirmed cases are only 10% of the actual cases, and that COVID-19 has an 0.75% infection-fatality rate (IFR), as we justify in the supplemental appendix, we would expect 300,000 to 600,000 deaths by the end of September 2020 if the social distancing occurs at 50% to 100% levels of normalcy, respectively. We note, however, based on our forecast that 44% of the counties in the sample (1,196) could completely reopen and still experience a confirmed case rate lower than 0.1% from June to August, 2020. These counties are less populated and account for less than 15% of the population in our sample. We also note that our analyses do not consider the positive effects of alternative preventive protocols such as facial masks wearing and hand sanitizing. Such protocols may help

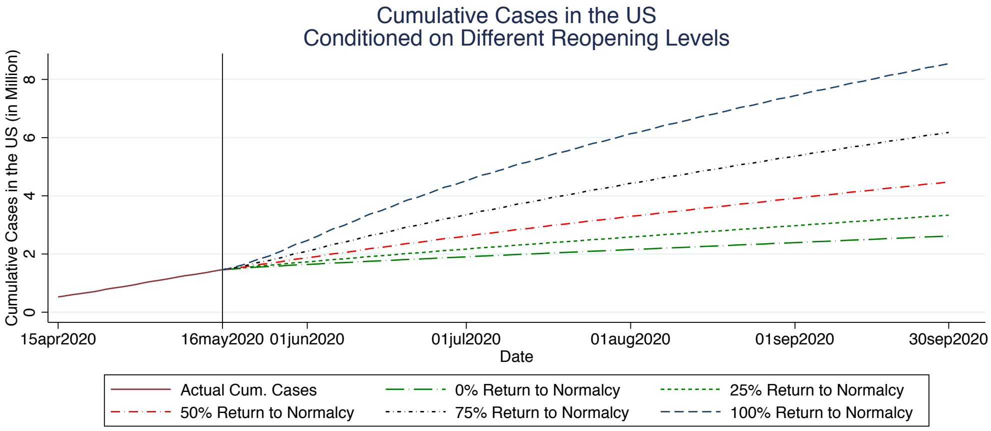

Figure 2: Cumulative Case Forecasting under Different Reopening Strategies. The vertical line indicates the last day of case data sample.

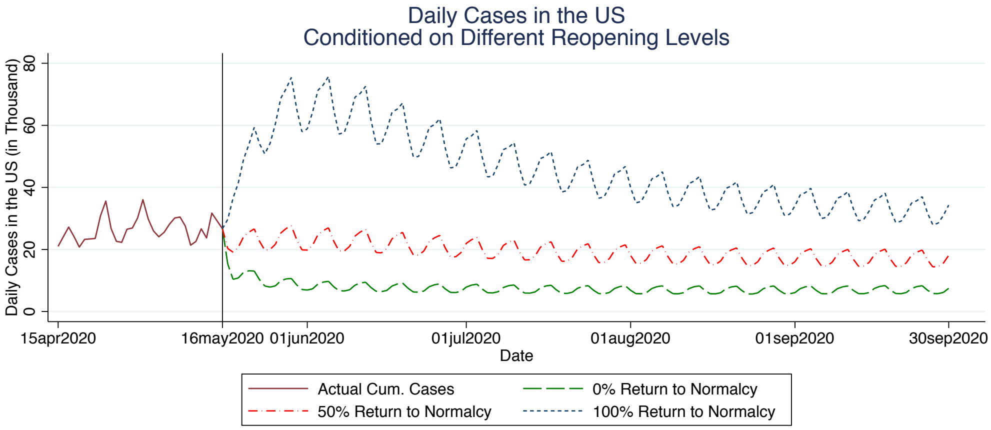

Figure 3: Daily Case Forecasting under Different Reopening Strategies. The vertical line indicates the last day of case data sample.

slow the contagion process.

Taken together, we demonstrate that the rate of spread of COVID-19 in the United States is concave in the number of contagious individuals. This explains why the growth rate of COVID-19 cases has been slower than expected given the initial exponential growth, above and beyond the effect from social distancing. We empirically identify the causal and substantial impact of social-distancing on combating the pandemic. We also forecast how COVID-19 will evolve in the future, and the timing over which different parts of the country will reach their peaks and how the patterns may affect our reopening strategies.

References

Danis, K., O. Epaulard, T. Benet, A. Gaymard, S. Campoy, E. Bothelo-Nevers, M. Bouscambert-Duchamp, G. Spaccaferri, F. Ader, A. Mailles, Z. Boudalaa, V. Tolsma, J. Berra, S. Vaux, E. Forestier, C. Landelle, E. Fougere, A. Thabuis, P. Berthelot, R. Veil, D. Levy-Bruhl, C. Chidiac, B. Lina, B. Coignard, C. Saura, and I. Team (2020): "Cluster of coronavirus disease 2019 (Covid-19) in the French Alps," *Clinical Infectious Diseases.*

Hansen, C. B. (2007): "Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects," *Journal of Econometrics*, 140, 670 – 694.

Nishiuram, H., N. M. Linton, and A. R. Akhmetzhanov (2020): "Serial interval of novel coronavirus (COVID-19) infections," *International Journal of Infectious Diseases*, 93, 284–286.

Oliveiros, B., L. Caramelo, N. C. Ferreira, and F. Caramelo (2020): "Role of temperature and humidity in the modulation of the doubling time of COVID-19 cases," *Working Paper*, accessed on May 20, 2020 at https://www.medrxiv.org/content/10.1101/2020.03.05.20031872v1.

Wang, J., K. Tang, K. Feng, and W. Lv (2020a): "High Temperature and High Humidity Reduce the Transmission of COVID-19," *Working Paper*, accessed on May 20, 2020 at SSRN: https://ssrn.com/abstract=3551767 or http://dx.doi.org/10.2139/ssrn.3551767.

Wang, M., A. Jiang, L. Gong, L. Luo, W. Guo, C. Li, J. Zheng, C. Li, B. Yang, J. Zeng, Y. Chen, K. Zheng, and H. Li (2020b): "Temperature significant change COVID-19 Transmission in 429 cities," *Working Paper*, accessed on May 20, 2020 at https://www.medrxiv.org/content/10.1101/2020.02.22.20025791v1.