Forecasting the Spread of COVID-19 under Different Reopening Strategies

Meng Liu[1,*,+], **Raphael Thomadsen**[1,*,+], **and Song Yao**[1,*,+]

[1]Olin Business School, Washington University in St. Louis, MO, 63130, U.S.A.
[*]Corresponding author emails: mengl@wustl.edu, thomadsen@wustl.edu, songyao@wustl.edu
[+]These authors contributed equally to this work and are listed alphabetically

ABSTRACT

We combine COVID-19 case data with demographic and mobility data to estimate a modified susceptible-infected-recovered (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 large effects on the rate of contagion. The social distancing in late March and April substantially reduced the number of COVID-19 cases. However, the concave contagion pattern means that when social distancing measures are lifted, the growth rate is considerable but will not be exponential as predicted by standard SIR models. Furthermore, counties with the lowest population density could likely avoid high levels of contagion even with no social distancing. We forecast rates of new cases for COVID-19 under different social distancing norms and find that if social distancing is eliminated there will be a massive increase in the cases of COVID-19, about double what would occur if the US only restored to 50% of the way to normalcy.

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 had 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 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 number of contagious individuals. We also observe that social distancing during late March and April significantly reduced the spread of the disease. Higher population density also leads to an increased spread of COVID-19.

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 country continues the nearly 50% return to normalcy observed at the end of April (as compared to the observed peak social distancing levels). 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 than we would expect if the country continued with the path of 50% return to normalcy we observe at the end of April. 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 of less than 0.1% over a 3-month period. These counties all have low population densities.

Model

The model we estimate is a simplified version of a susceptible-infected-recovered (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 by 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[1]. This treatment of the infectious population is an approximation to the standard SIR models, where the infectious population is typically 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 of 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 $(Y_{i,t-2} - Y_{i,t-16})$ 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 constraint 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 acquaintances. 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[2],[3],[4]. The social distancing measure, $d_{i,t}$, 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, underestimated impact of social distancing on slowing the spread of the disease. We control for this endogeneity bias by estimating the model using an Instrumental Variables (IV) 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 $\varepsilon_{i,t}$.

Results

To estimate 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 omit the data from New York, New Jersey and Connecticut, as explained in the technical appendix. We append the data with daily county-level weather data as well as cellphone mobility data provided by SafeGraph. The results are presented in Table 1. 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%.

We also find that the exponent on the number of contagious people, ω, is 0.47. It is 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 at which the disease spreads. This level of concavity also implies that while initial outbreaks of COVID-19 will expand exponentially, they will quickly turn to a slower rate of growth. The growth looks linear or even plateauing when 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 recent growth rate of COVID-19 cases has slowed considerably after a quick take-off, and yet this growth has persisted. We also find that 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 the county fixed effects α_i against county-level demographics. It has been shown in [5] that when the number of fixed effects is large, these coefficients can be treated as data for the purpose of statistical inference. The results are reported in Table 2. We find that population density is a crucial factor influencing the spread of the disease. In fact, COVID-19 would be expected to never flare up beyond a very small base level in

Dependent Variable	Log(Infected in County *i* on Date *t*)
Social Dist. Level in	-5.445***
County *i* on Date *t*	(1.656)
Infectious Individuals in	0.470***
County *i* on Date *t*	(0.014)
Avg. Temperature (°C) of	-0.009***
County *i* on Date *t*	(0.002)
County Fixed Effects	Yes (Mean=0.262)
Date Fixed Effects	Yes (Mean=3.835)
Observations	73,713
R_squared	0.614
Counties	2,704

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

Table 1. Estimation of a Modified SIR Model.

Dependent Variable	County Fixed Effect
Log(Pop. Density)	0.393***
(People/Sq. Miles)	(0.010)
Fraction of Black	1.418***
Residents	(0.101)
Fraction of Hispanic	1.376***
Residents	(0.105)
Fraction of Commuters	5.444***
using Pub. Transportation	(0.664)
Log(Median Income)	1.124***
(in U.S. dollars)	(0.063)
Fraction of Senior	1.047*
Residents (≥70yrs)	(0.556)
Fraction of Children	0.780
Residents (<18yrs)	(0.509)
Constant	-14.111***
	(0.692)
R_squared	0.619
Counties	2,703

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

Table 2. Factors Affecting County Fixed Effects of Rate of Contagion

some areas with sufficiently low population densities, as we discuss below. 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 greater contagion. We are uncertain what drives this result: one possibility is that people with higher incomes may interact more with nearby cities that have more outbreaks. We also include the share of seniors (age 70+) and children (below age 18) in the population. Seniors are marginally more likely to spread the disease, but children show no sign of having a lower rate of infecting people, confirming the finding in [6].

Prediction and Forecasting

Using our model, we simulate future cases beyond our sample period. 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% between the peak lock-down level and normalcy (so we say that such a level is at 50% of normalcy). 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 30, 2020 under different reopening strategies. The cumulative and daily cases appear in Figures 2 and 3, respectively. The cyclicality 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 to 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 to a full return to normalcy will lead to 4 million additional confirmed cases. If we assume that confirmed cases are only 10% of 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 wearing facial masks and better hand washing. Such protocols may help slow the contagion process.

Conclusion

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 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

1. Nishiuram, H., Linton, N. M. & Akhmetzhanov, A. R. Serial interval of novel coronavirus (covid-19) infections. *Int. J. Infect. Dis.* **93**, 284–286 (2020).

2. Wang, J., Tang, K., Feng, K. & Lv, W. High temperature and high humidity reduce the transmission of covid-19. *Work. Pap.* (2020). Accessed on May 20, 2020 at SSRN: https://ssrn.com/abstract=3551767 or http://dx.doi.org/10.2139/ssrn.3551767.

3. Oliveiros, B., Caramelo, L., Ferreira, N. C. & Caramelo, F. Role of temperature and humidity in the modulation of the doubling time of covid-19 cases. *Work. Pap.* (2020). Accessed on May 20, 2020 at https://www.medrxiv.org/content/10.1101/2020.03.05.20031872v1.

4. Wang, M. *et al.* Temperature significant change covid-19 transmission in 429 cities. *Work. Pap.* (2020). Accessed on May 20, 2020 at https://www.medrxiv.org/content/10.1101/2020.02.22.20025791v1.

5. Hansen, C. B. Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *J. Econom.* **140**, 670 – 694, DOI: https://doi.org/10.1016/j.jeconom.2006.07.011 (2007).

6. Danis, K. *et al.* Cluster of coronavirus disease 2019 (covid-19) in the french alps. *Clin. Infect. Dis.* DOI: 10.1093/cid/ciaa424 (2020).

Acknowledgments

We thank SafeGraph Inc. for providing the mobility data analyzed in this study.

Author contributions statement

All authors formulated the empirical model. M.L. and S.Y. analyzed the data. All authors drafted and reviewed the manuscript.

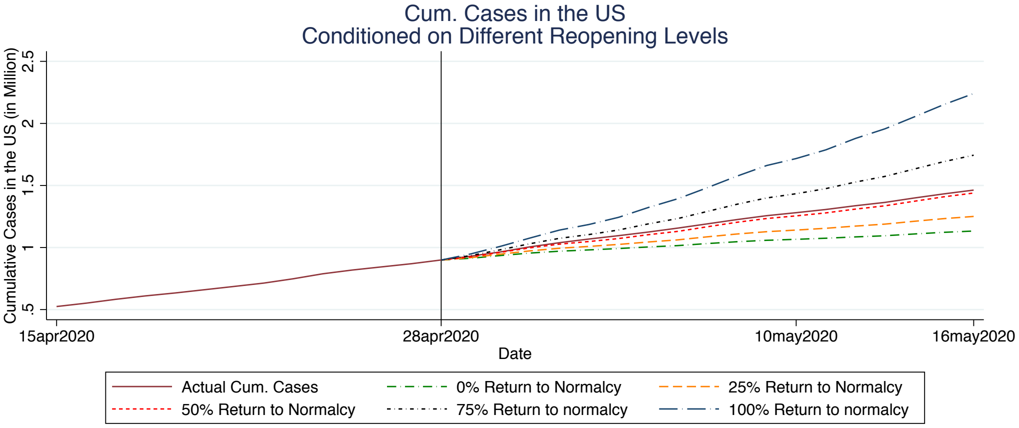

Figure 1. Out-of-Sample Prediction under Different Reopening Strategies. The vertical line indicates the last day of our data sample.

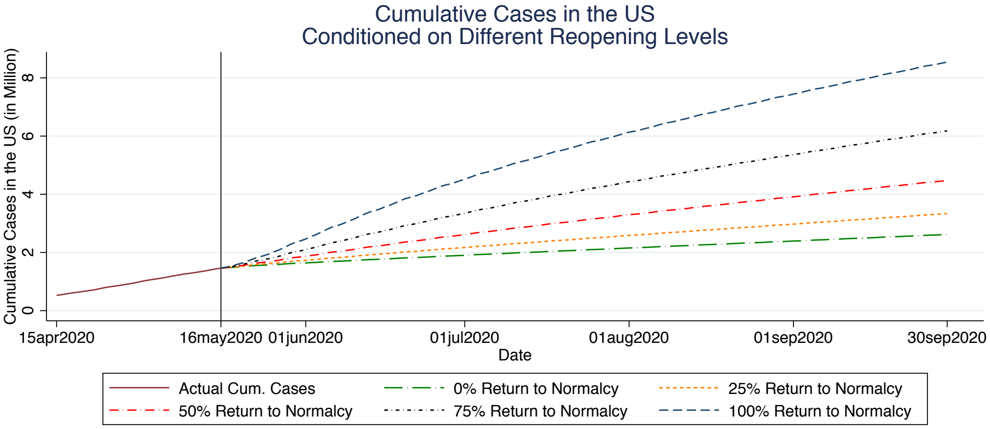

Figure 2. Cumulative Case Forecasting under Different Reopening Strategies. The vertical line indicates the last day of case data sample.

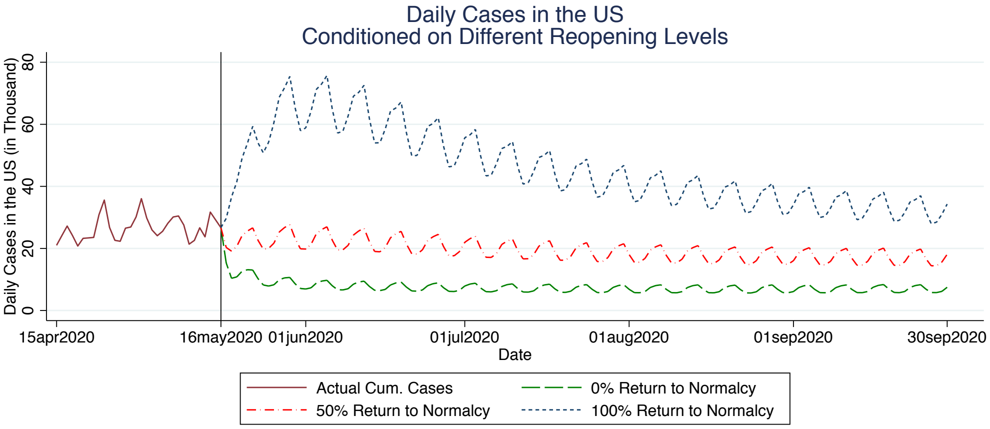

Figure 3. Daily Case Forecasting under Different Reopening Strategies. The vertical line indicates the last day of case data sample.

Technical Appendix: Forecasting the Spread of COVID-19 under Different Reopening Strategies

Meng Liu[1,*,+], **Raphael Thomadsen**[1,*,+], **and Song Yao**[1,*,+]

[1]Olin Business School, Washington University in St. Louis, MO, 63130, U.S.A.
[*]Corresponding author emails: mengl@wustl.edu, thomadsen@wustl.edu, songyao@wustl.edu
[+]These authors contributed equally to this work and are listed alphabetically

Introduction

In this online appendix, we first present our data. We then discuss our model and assumptions. Finally, we present our assumptions for the simulations.

Data

Our data come from a multitude of sources. We lay out the sources for each of these in turn.

Positive Cases

Data of positive cases are based on the COVID-19 data published by the New York Times (https://github.com/nytimes/covid-19-data, accessed on May 17, 2020). The data contain the daily confirmed case counts in the U.S. at the county or county-equivalents level. We exclude cases in the states of New York, New Jersey, and Connecticut due to the large outbreak there and the complicated relationship between New York City (which is the seat of 5 counties) and the surrounding counties. We drop 3 of the remaining counties because we do not have social distancing data for 2 of them, and we cannot match the demographic data for a third (Oglala Lakota County, SD). This reduces the number of counties to 2,778. In addition, we remove 74 counties that had no confirmed cases in the entire sample period. The remaining 2,704 counties and county-equivalents constitute our main sample. These counties account for 89.6% of the total U.S. population and 55.2% of the total U.S. confirmed cases as of April 23, 2020.

There are a few days where there are negative cases that are reported. These are generally corrections to previous over-reporting. Thus, we clean the negative numbers of cases by subtracting the absolute value of the negative cases from the proceeding day. In the event that that leads to a negative number of the proceeding day, we iterate again.

COVID-19 also has an incubation period of approximately 5 days [1],[2]. Because of this lag between when a person gets sick and when they are diagnosed with COVID-19, we assume that the cases reported on a particular date actually measure the COVID-19 infections from 5 days earlier. We also assume that the true number of cases is approximately 10 times the number of diagnosed cases. We get this number by assuming that the Infection Fatality Rate (IFR) is 0.75% [3]. We also assume that any deaths that occur happen 14 days after the confirmed test result. On May 16, 2020, the last day of our confirmed case data, there were 88,660 deaths in the US. On May 2, 2020, there were 1,138,961 officially diagnosed cases. We hence obtain the factor as (88,660/0.0075)/1,138,961 = 10.4. We round this number to 10. Our estimates are not sensitive to the specific factor we use.

Social Distancing

We use social distancing data from the company SafeGraph, which collects cellphone GPS data from U.S. residents, and has made them available for free to academics studying COVID-19. These data are collected through a series of pings that the company receives for all users who have installed a number of smartphone apps. The list of apps that collect this information is kept as a trade secret. For each county, we use the fraction of cellphones that stayed near home for the whole day as our measure of social distancing. The SafeGraph data are published at the Census Block Group level. To accommodate other data sources which are available at a less granular level, we aggregate the this variable to the county level by taking the weighted median, using the number of cellphones in each Census Block Group as the weight.

Demographic data

We obtain the demographic data from the Census Bureau's 2014-2018 American Community Survey (ACS), which contains information of each county's profile of population, ethnicity, age, median income, and commuting pattern. The ACS, however, does not report population densities. SafeGraph, the company who provides us with the social distancing data, also maintains a dataset of the land area of each Census Block Group in the US. We aggregate the land areas to the county level. Together with the county population information from the Census Bureau, we are able to construct the population density data of each county.

Weather data

We gathered historical daily rain and temperature data from National Oceanic and Atmospheric Administration (NOAA) (source: https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00861/html, accessed on May 21, 2020). The raw weather data is at the weather station level and we match weather stations to the counties they are in. We use the average values across weather stations within the same county to construct the weather variables for that county. For a small number of counties where there are no associated weather stations, we use the daily state averages as proxies.

Voting data

We obtained county-level voting data from https://public.opendatasoft.com/explore/dataset/usa-2016-presidential-election-by-county/table/?disjunctive.state (accessed May 20, 2020). This data is explained at https://github.com/Deleetdk/USA.county.data (accessed May 20, 2020), and the election vote totals originally came from the New York Times.

Shelter-in-place orders data

Shelter-in-place orders (SIP) data are compiled by Keystone, a strategy and economics consulting firm. The company collects and distributes the SIP data (https://www.keystonestrategy.com/coronavirus-covid19-intervention-dataset-model/, accessed on May 21, 2020.) for free to researchers studying COVID-19.

Putting it all together

Our sample is an unbalanced panel because counties start to have positive number of confirmed cases on different dates. The earliest date we observe in the sample is Jan 29, 2020, and the last day is April 23, 2020. Note that we construct actual cases using reported cases 5 days later, and thus the corresponding sample period based on reported cases is Feb 3, 2020 to April 28, 2020.

Summary statistics of all of the variables we use in the estimation are presented in Table A1. Note that our case data proceed past the dates used for estimating the model and are up to May 16, 2020. We use those data for validating the model. Those data are publicly available, but we are happy to supply summary statistics for this hold-out sample upon request.

Empirical Analysis

Detail of Empirical Specification

In this subsection, we detail the assumptions we make and the estimation procedure. As noted in the main paper, the model we estimate is a modified version of the standard susceptible-infected-recovered (SIR) model:

$$y_{i,t} = R_{i,t} S_{i,t} \left(Y_{i,t-2} - Y_{i,t-8} \right)^{\omega} \tag{A1}$$

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 population that has not yet had COVID-19 and is thus susceptible to it, and $Y_{i,t}$ is the number of cumulative individuals who have been infected up until day t. This model differs from the standard SIR in two key ways. First, the standard SIR model constrains $\omega = 1$. We discuss in the paper that there are theoretical reasons to believe that the correct model of transmission involves $\omega < 1$. As an example, we present a network model below that demonstrates that $\omega < 1$ is possible even using the conventional transmission mechanism. Second, the standard SIR model does not specify a discrete time frame over which the infected individuals are contagious, but rather builds a stock of infected individuals and assumes that these individuals exit their infected period at a fixed rate. We view our model as an approximation of this process, which greatly eases our estimation and allows us to easily add important variables to explain the contagion process in our analysis. We use a 6-day infectious window and an assumed latent period of 2 days $\left(Y_{i,t-2} - Y_{i,t-8}\right)$. This gives a mean serial interval of 4.5 days, which is close to several estimates [4]. In this appendix, we also present the results where we use a 14-day window $\left(Y_{i,t-2} - Y_{i,t-16}\right)$, and show that results are similar.

We assume that the rate of transmission, $R_{i,t}$ varies by a set of factors, which we model as $R_{i,t} = \exp\left(X'_{i,t}\beta + \varepsilon_{i,t}\right)$. $X_{i,t}$ includes county-level fixed effects, date fixed effects, the measure of social distancing, and daily average temperature. The county fixed effects account for differences in demographics across counties, such as the demographics shown in Table 2 of the main paper as well as other unobservable county-specific factors. The date fixed effects account for both day-of-the-week differences in the patterns of travel for people (e.g., the time away from the house to go to work or to go to the park, which may lead to different exposures to the disease) as well as differences in the rate of testing and reporting that occur across time. We assume that the errors $\varepsilon_{i,t}$ are uncorrelated across counties. We further assume that $\varepsilon_{i,t}$ is uncorrelated across time, although we cluster the standard errors by county.

We estimate the model by taking logarithm of both sides. After rearranging we get:

$$\left[\ln\left(y_{i,t}\right) - \ln\left(S_{i,t}\right)\right] = X'_{i,t}\beta + \omega \ln\left(Y_{i,t-2} - Y_{i,t-8}\right) + \varepsilon_{i,t} \tag{A2}$$

Note that sometimes $y_{i.t}$, the diagnosed case number, is 0 for some counties on some dates. Therefore, we adjust this formula slightly by adding 1 to $y_{i,t}$ so the logarithmic values are always well-defined:

$$\left[\ln\left(y_{i,t} + 1\right) - \ln\left(S_{i,t}\right)\right] = X'_{i,t}\beta + \omega \ln\left(Y_{i,t-2} - Y_{i,t-8}\right) + \varepsilon_{i,t} \tag{A3}$$

In some counties, $Y_{i,t-2} - Y_{i,t-8}$ is 0 for some periods. We do not use those observations for estimation. Note that because this is a lagged variable, this is a selection based on independent variables and not based on dependent variables, and hence it does not bias our estimation.

One concern that can arise in estimating this model is that the amount of social distancing is likely to be correlated with the error terms, $\varepsilon_{i,t}$ in the regression. We address this concern using an Instrumental Variable (IV) approach, which requires that we find a variable that affects social distancing but is not correlated with $\varepsilon_{i,t}$. We use the amount of rain (measured in mm) as a shifter of social distancing that does not directly cause COVID-19 to spread. We run a first stage regression of social distancing to test whether this instrument has much power. The F-statistic for this test is 435.43, indicating that this is a strong instrument. The main estimation results are presented in Table 1 of the paper and replicated in column 1 of Table A2 in this appendix.

Research on COVID-19 is nascent, and there are different views of how long infected individuals stay contagious. Suppose that such individuals are contagious for 14 days instead of 6 days. Then the model becomes:

$$y_{i,t} = R_{i,t} S_{i,t} \left(Y_{i,t-2} - Y_{i,t-16}\right)^{\omega}. \tag{A4}$$

We present the estimation results of this model in column 2 of Table A2. Note that this regression has more observations because there are fewer instances where we observe no cases in a county for a 14-day window than for a 6-day window. The results are largely unchanged. The coefficient on social distancing levels are slightly lower, but well within one standard error of the corresponding coefficient in column 1. The exponent on the infectious individuals is 0.435. That is somewhat smaller (but statistically different) than the 0.470 we observe with the shorter 6-day window, but overall the curvature shape is similar to what we have observed with the 6-day window. The effect of temperature is slightly smaller but similar. As we will discussion below in Section **Simulation**, both specifications give similar long-run forecasting results.

All county-level demographic factors remain constant over time in our analysis. While our main regression gives many insights, impacts of these demographic factors on the spread of the virus are captured by the county fixed effects. In order to better understand how these factors affect the contagion rate, we next regress the county fixed effects on several demographic variables. The coefficients from this regression should be thought of as the impact of these demographics on the rate of contagion. The results from the model are reported in column 1 of Table A3 (replication of Table 2 in the paper). In column 2 we present the results we would obtain if we instead modeled the contagious period to be 14 days. Similar to Table A2, the results are again very similar under this alternative specification. The only statistically significant difference is that the coefficient on log(population density) is slightly smaller, although the effect is of a very similar magnitude. We also observe that the R^2 of the 14-day contagious period model is slightly lower than the R^2 of the 6-day contagious period model.

Simulation

After estimating our model, we forecast the number of cases that would emerge under different social distancing regimes. For this exercise, we first divide our model's predicted numbers of "true" cases by 10, which gives us the prediction of diagnosed cases (as described in Section **Positive Cases**). Next, because the 2,704 counties in our sample are a subset of the whole nation, on a given date the predicted diagnosed case number is a fraction of the total cases in the US. While this fraction changes on a daily basis, we use an approximation by taking the median of the daily ratios between diagnosed cases observed in our sample and in the whole nation during our sample period. The median of the daily ratios is 0.63. Accordingly, we divide our predicted diagnosed cases by 0.63 to obtain the national number of diagnosed cases.

The first step of our simulation involves validating the model: We predict how many cases would emerge in the weeks after our data (Dates: April 29 to May 16, 2020) in order to validate our model. The results are shown in Figure 1 in the original paper. We observe that we are well able to predict the number of observed cases if the social distancing in early May represented a 50% return to normalcy, which is defined as $\frac{FractionStayAtHomePeak_i - FractionStayAtHome_{i,t}}{FractionStayAtHomePeak_i - FractionStayAtHomeBeforeSD_i}$. $FractionStayAtHomePeak_i$ represents the fraction of devices staying home at the peak of social distancing in our data (April 5-April 11, 2020). This variable is county-specific, hence the county subscript i. $FractionStayAtHomeBeforeSD_i$ represents the observed lowest level at which devices stayed home in county i in February, and $FractionStayAtHome_{i,t}$ represents the fraction of devices staying home in county i on date t. We compute the different levels of social distancing accordingly: For example, a 25% towards normalcy represents social distancing at the level of 0.25×(minimum social distancing) + 0.75×(maximum social distancing). We find that the levels of social distancing at the end of April were approximately at the 50% return to normalcy levels. Overall, our model predicts the national cases well.

We next forecast the cumulative and daily cases of COVID-19 through the end of September at different levels of social distancing. Those forecasts appear in Figures 2 and 3 in the original paper. In Figure A1 below, we replicate Figure 2 in the paper but further add the confidence intervals. To avoid cluttering, we only depict 50% and 100% return-to-normalcy levels in A1.

As a robustness check regarding the 6-day contagion window specification, we also consider forecasting US daily cases under the specification where the contagion window is 14 days. Figure A2 shows the evolutions of daily cases till September 30, 2020 under 50% and 100% return-to-normalcy regimes. We overlay the forecasts of both 14-day and 6-day specifications for easy comparison. From the figure, we may see the forecasts of 14-day and 6-day contagion window specifications are fairly close under the 50% return-to-normalcy level. For the 100% normalcy level, the two specifications differ at the beginning but converge quickly. In the long run, the two specifications give similar forecasts for daily cases.

Concavity of SIR model and Network Dynamics

One unique feature of our model is that we estimate an exponent on the number of contagious cases. We include this flexibility because such a model fits the data much better, and also leads to forecasts that have more limited growth after an initial take-off of COVID-19 cases, as is commonly observed. To demonstrate our model's better fit, we compare the prediction of our model against an alternative model with the exponent fixed at 1 as the standard SIR model in Figure A3. At the 50%-return-to-normalcy level that is observed at the end of April, the standard SIR model would predict much higher cumulative case numbers than the actual numbers. In contrast, our model's prediction fits much better.

We next illustrate that the concave relationship we estimate for the number of contagious individuals on the number of new cases can come from social networks between people. We seek to demonstrate the theoretical feasibility of our results rather than the necessary or sufficient conditions under which the nonlinearity will arise. Thus, we simulate a very simplified model of networks and disease process.

To do this, we simulate a network with the following process: We take 10,000 individuals. We create a network by first randomly assigning that any two individuals will be joined with a common node with probability 0.001999 (corresponding to each person getting almost 20 friends on average). Call these connection "round-1 friends." We then expand this network by assigning each node to have an edge with each of the round-1 friends of their friends with a probability of 0.6.

We assume that the disease spreads with the following process. We seed 4 individuals to have the disease in period 0. Then in each period we assume that any connected individual will get sick with probability 7/(number of connections), where the number of connections is specific to the individual, and varies due to the random assignment of the people who are connected (This probability is capped at 1 in case someone is randomly assigned fewer than 7 connections, which is very unlikely).

After simulating this process, we then regress $(\ln(y_t) - \ln(S_t)) = c + \omega \ln(y_{t-1}) + \varepsilon_t$. We run this simulation 10 times. The mean value for $\hat{\omega} = 0.51$, with a range of (0.44, 0.62). This shows the plausibility of network effects leading to an estimate in the range that we have estimated in our main model.

References

1. Lauer, S. A. *et al.* The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals Intern. Medicine* **172**, 577–582 (2020).

2. Li, Q. *et al.* Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *The New Engl. J. Medicine* **382**, 1200–1207 (2020).

3. Meyerowitz-Katz, G. & Merone, L. A systematic review and meta-analysis of published research data on covid-19 infection-fatality rates. *Work. Pap.* (2020).

4. Nishiuram, H., Linton, N. M. & Akhmetzhanov, A. R. Serial interval of novel coronavirus (covid-19) infections. *Int. J. Infect. Dis.* **93**, 284–286 (2020).

	(1)	(2)	(3)	(4)	(5)
Time-varying Variables	N	Mean	Std Dev.	Min	Max
Reported cumulative cases	73,713	95.53	547.4	0	25,811
Reported new cases	73,713	6.437	34.62	0	1,475
Average temp (in Celsius)	73,713	11.35	7.144	-23.38	30.70
Rain (mm)	73,713	3.397	8.873	0	197.9
Social_distancing (home ratio)	73,713	0.349	0.0727	0.0870	0.828
Time-invariant Variables					
Median income ($)	2,703	52,293	14,291	19,943	144,821
Logged median income	2,703	10.83	0.259	9.901	11.88
Population	2,704	108,809	343,978	625	1.004e+07
Population density (people / square miles)	2,704	224.6	794.8	0.0359	18,799
Logged population density	2,704	4.016	1.567	-3.326	9.842
Share of population senior (>=70)	2,704	0.121	0.0325	0.0229	0.384
Share of population youth (<=17)	2,704	0.225	0.0336	0.0732	0.403
Share of population black	2,704	0.100	0.152	0	0.874
Share of population Hispanic	2,704	0.0915	0.135	0	0.991
Share of population public transit commuters	2,703	0.00844	0.0219	0	0.371
Fraction votes for Trump 2016	2,690	0.628	0.154	0.0412	0.953

Table A1. **Summary Statistics**

	(1) Contagious for 6 days	(2) Contagious for 14 days
Dependent Variable	Log(Infected in County *i* on Date *t*)	Log(Infected in County *i* on Date *t*)
Social Dist. Level in	-5.445***	-4.959***
County *i* on Date *t*	(1.656)	(1.490)
Infectious Individuals in	0.470***	0.435***
County *i* on Date *t*	(0.014)	(0.015)
Avg. Temperature $(^{\circ}C)$ of	-0.009***	-0.008***
County *i* on Date *t*	(0.002)	(0.002)
County Fixed Effects	Yes	Yes
Date Fixed Effects	Yes	Yes
Observations	73,713	83,393
R_squared	0.61	0.62
Counties	2,704	2,704

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

Table A2. **Estimation of a Modified SIR Model.**

	(1) Contagious for 6 days	(2) Contagious for 14 days
Dependent Variable	County Fixed Effect	County Fixed Effect
Log(Pop. Density)	0.3930***	0.3664***
(People/Sq. Miles)	(0.0102)	(0.0102)
Fraction Black Residents	1.4175***	1.3730***
	(0.1008)	(0.1008)
Percentage Hispanic	1.3755***	1.3609***
Residents	(0.1051)	(0.1052)
Percentage of Commuters	5.4436***	6.0386***
using Pub. Transportation	(0.6639)	(0.6642)
Log(Median Income)	1.1244***	1.0511***
(in U.S. dollars)	(0.0626)	(0.0626)
Percentage of Senior	1.0473*	1.4409***
Residents ($\geq$70yrs)	(0.5564)	(0.5566)
Percentage of Children	0.7803	1.2441**
Residents (<18yrs)	(0.5089)	(0.5091)
Constant	-14.1108***	-13.8588***
	(0.6923)	(0.6925)
R_squared	0.62	0.59
Counties	2,703	2,703

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

Table A3. Analysis of County Fixed Effects

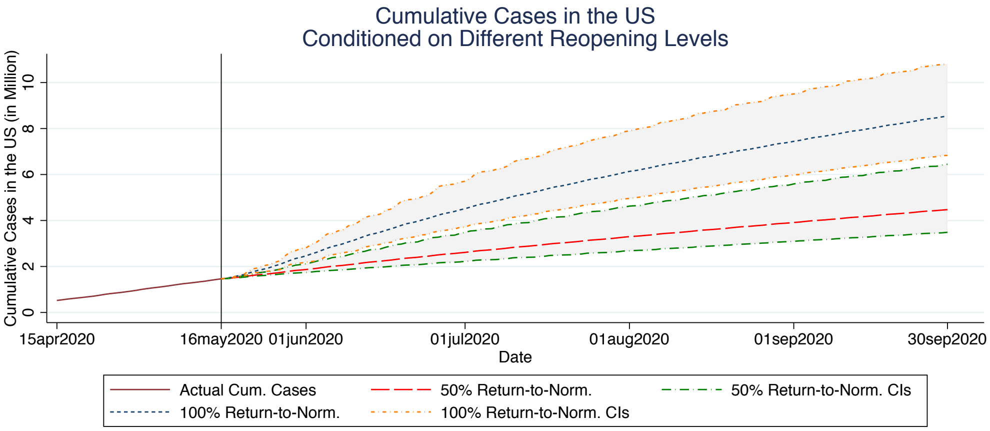

Figure A1. Cumulative Case Forecasting under Different Reopening Strategies with Confidence Intervals. The vertical line indicates the last day of case data sample.

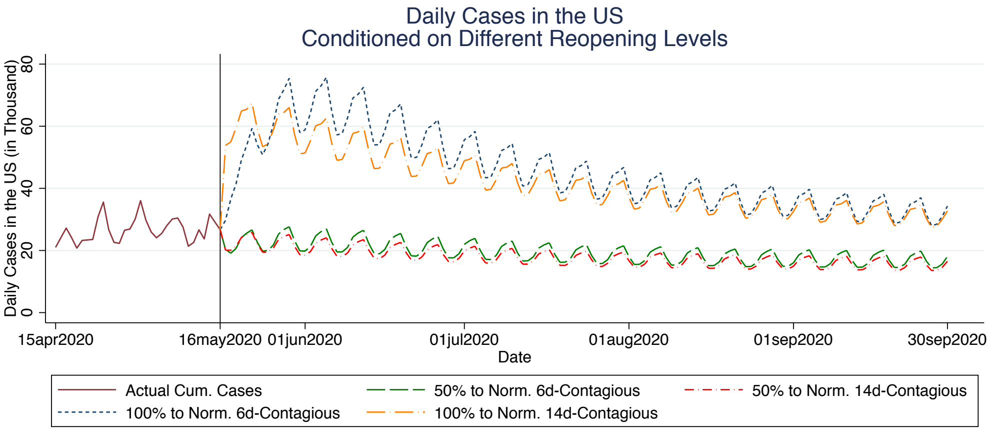

Figure A2. Daily Case Forecasting under Different Contagion Window Specifications. The vertical line indicates the last day of case data sample.

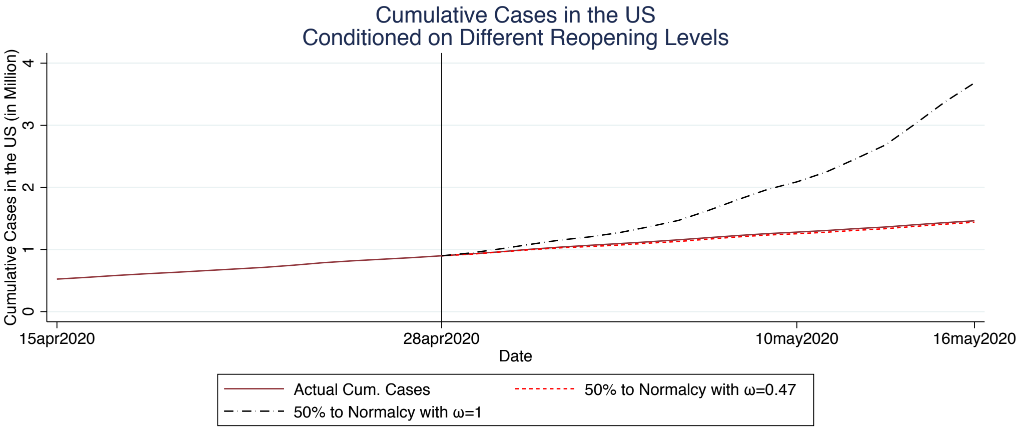

Figure A3. Out-of-Sample Prediction with the Standard SIR Model. The vertical line indicates the last day of our data sample.