scientific reports

Check for updates

OPEN

Forecasting the spread of COVID-19 under different reopening strategies

Meng Liu[1,2], Raphael Thomadsen[1,2] & Song Yao[1,2]✉

We combine COVID-19 case data with mobility data to estimate a modified susceptible-infected-recovered (SIR) model in the United States. In contrast to a standard SIR model, we find that the incidence of COVID-19 spread is concave in the number of infectious individuals, as would be expected if people have inter-related social networks. This concave shape has a significant impact on forecasted COVID-19 cases. In particular, our model forecasts that the number of COVID-19 cases would only have an exponential growth for a brief period at the beginning of the contagion event or right after a reopening, but would quickly settle into a prolonged period of time with stable, slightly declining levels of disease spread. This pattern is consistent with observed levels of COVID-19 cases in the US, but inconsistent with standard SIR modeling. 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.

The COVID-19 pandemic has caused great disruption. Over 43 million people have confirmed diagnoses of the disease, and over 1 million people have died from it[1]. It has also had substantial impacts on daily lives and economic activities[2,3]. Many studies have focused on measuring who are affected the most by COVID-19[4,5], or which therapies are appropriate at each stage of the disease[6–8]. However, it is also crucial to understand how the spread of COVID-19 depends on preventive measures such as social distancing and how the reopening may affect the spread.

The most common model used to study the spread of COVID-19 is the susceptible-infected-recovered (SIR) model. In such models, there is a susceptible population, which is assumed to be equal to the population of whichever region is being examined minus the number of people that have previously had the disease. Some of the susceptible individuals get infected in each period, where the rate of infection is a function of the number of infectious individuals as well as other factors that shift the rate of transmission. Finally, infectious individuals move to a state of recovery. In our analysis, we call anyone who was sick but is no longer infectious to be "recovered," although some of these people may still actually be sick, hospitalized, or have died. Thus, the recovered terminology is actually a shorthand for all post-infectious states. This model, and its variants, have been used extensively to study the growth of COVID-19. For example, a recent study estimates a Susceptible-Exposed-Infected-Confirmed-Removed (SEIQR) model, which appends the standard SIR model with a stage modeling susceptible people who become exposed to the virus and a stage modeling infected people which are confirmed to have the disease[9]. The paper then applies this model to estimate COVID-19 transmission in Wuhan, China, showing that an earlier lockdown makes the outbreak worse in Wuhan but helps the rest of the world. The SEIQR model is also used to show that travel restrictions may have reduced the spread of COVID-19 from Wuhan, China, to other Chinese cities[10,11]. As another variant to the SIR model, the SEIR model (adding an exposure stage to the SIR model) is also applied to compute the rate of transmission both from animals to people and from people to people[12]. While it may seem that having more stages in the model would make the SEIR model superior to the SIR model, it has been shown that the standard SIR model does a better job at predicting the spread of COVID-19, based on data from Wuhan, China[13].

In this paper, we use a modified version of the SIR model to measure the extent to which social distancing reduces the speed at which COVID-19 spreads. We then run simulations to forecast the rates of COVID-19 spread under different social distancing levels during the reopening. We find that COVID-19 spreads less than proportionately with the number of infectious individuals, a distinct difference from the assumption of standard models. We demonstrate that this pattern could be explained by the interconnectedness of people's social networks. This pattern suggests that each additional infectious individual has less impact on the disease spread

[1]Olin Business School, Washington University in St. Louis, Missouri 63130, USA. [2]These authors contributed equally: Meng Liu, Raphael Thomadsen and Song Yao. ✉email: songyao@wustl.edu

as more people become infected. One key implication of this finding is that the rate of disease growth can be slow and steady, rather than either exponential or falling quickly, as would be implied by the most-commonly used models. This leads to more accurate predictions of the spread of COVID-19. We also observe that social distancing greatly reduces the spread of COVID-19.

Mathematically, we model transmission of COVID-19 as

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

where $y_{i,t}$ is the number of new infections in county i on date t, $R_{i,t}$ is the rate at which infectious individuals transmit the disease, $S_{i,t}$ is the percentage of the county population that has not yet had COVID-19, and $Y_{i,t}$ is the cumulative number of individuals who have been infected by date t. Correspondingly, the $Y_{i,t-2} - Y_{i,t-8}$ term reflects our assumption that infected individuals are infectious from the second day after they catch the virus through the seventh day. This implies that the average serial interval is 4.5 days under the assumption that the level of infectiousness and the level of contact with susceptible individuals is constant during this time[14]. This treatment of the infectious population is an approximation of the standard SIR model, where the infectious population is typically modeled as a stock that has a constant outflow rate. Discretizing the rate of transmission enables the estimation of a large number of county and date fixed effects in our model, and as a practical matter this assumption has little impact on our estimates of the contagion rate. As a robustness check, we obtain extremely similar COVID-19 forecasts if we take the time of infectiousness to be 14 days, $Y_{i,t-2} - Y_{i,t-16}$, instead of 6 days, as presented in the appendix. The main difference between our model and the standard SIR model is the inclusion of the exponent ω on the number of infectious individuals. This ω allows the rate of growth of COVID-19 to be less than proportionate with the number of infectious individuals if $\omega < 1$. Such a result would be expected if infectious individuals expose many of the same unexposed individuals, which could occur if people have overlapping social connections. We see this directly when, for example, cases are clustered within households, nursing homes, or places of work. Thus, we can think of ω as measuring the extent to which people's networks are more interconnected to a tight-knit group of individuals relative to their level of connectedness to the population as a whole.

We also allow the transmission rate $R_{i,t}$ to vary with a number of factors instead of treating it as a constant parameter:

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

We use $d_{i,t}$, $m_{i,t}$, and $h_{i,t}$ to represent the level of social distancing, temperature, and humidity in county i on date t, respectively, and $\varepsilon_{i,t}$ is the statistical error term. The parameters α and β are vectors of county and date fixed effects, where the i-th element of α, α_i, represents the fixed effect for county i. Similarly, the t-th element of β, β_t, represents the fixed effect for date t. These fixed effects measure the baseline transmission rate of each county and each date, respectively. The parameters λ, μ, and θ measure the impacts of social distancing, temperatures, and humidity on transmission rates, respectively. In short, this specification allows transmission rates to differ across counties (through the county fixed effects), dates (through the date fixed effects), levels of social distancing, temperatures, and humidity. We note that the impacts of the last two factors have been debated in the literature[15–17]. The county fixed effects account for differences in demographics across counties, such as the demographics shown in Table 2 below 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. As a robustness check, we also include the state-level testing numbers directly into Eq. (2) during estimation. The results are statistically indistinguishable from the main results, as noted in "Results and simulation" below.

The social distancing measure, $d_{i,t}$, is based on cellphone GPS location data that are provided by SafeGraph for free to researchers studying COVID-19. We measure social distancing as the first principle component of several daily measures of each county: the percentage of residents staying home, the percentage of residents working at workplace full-time, the percentage of residents working at workplace part-time, the median duration of residents staying home, and the median distance of residents traveled.

As noted earlier, the most crucial difference between our model and a standard SIR model is that a standard SIR model constrains the exponent $\omega = 1$. We instead find that $\omega = 0.57$. Thus, the marginal impact of one more infected person diminishes as more people are infected. Such a result would be expected if infectious individuals expose many of the same unexposed individuals within a clustered network of individuals. In the appendix we demonstrate that a networking model with contagion can yield $\omega < 1$.

Results and simulation

The estimated model appears in Table 1, with standard errors (s.e.) reported in the parentheses. The estimated exponent on the number of infectious people, ω, is 0.57. Thus, the number of new infections is concave with respect to the number of infectious individuals. This level of concavity also implies that while initial outbreaks of COVID-19 expand exponentially, the daily number of new cases quickly stabilizes to a long-term plateau. We also find that social distancing has a large impact on the growth rate of COVID-19, while humidity has a smaller effect and temperature is insignificant. (When including daily testing numbers of each state in Eq. (2), the estimates of ω and social distancing are 0.568 (s.e. = 0.014) and − 0.816 (s.e. = 0.246), respectively).

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

Dependent variable	Log(Infected in County *i* on Date *t*)
λ: Impact of Social Dist. Level in County *i* on Date *t*	− 0.824*** (0.245)
ω: Impact of Infectious Individuals in County *i* on Date *t*	0.571*** (0.014)
μ: Impact of Avg. Temperature (Celsius) of County *i* on Date *t*	− 0.001 (0.002)
θ: Impact of Avg. Humidity of County *i* on Date *t*	0.005** (0.002)
α: 2923 county fixed effects (Autauga County, AL is omitted as the baseline)	Estimated
β: 101 date fixed effects	Estimated
Observations	131,272
R_squared	0.63
Counties	2,924

Table 1. Estimation of a modified SIR model. *** p < 0.01, ** p < 0.05, * p < 0.1.

Dependent variable	County fixed effect of each county
Log(Pop. Density of Each County) (People/Sq. Miles)	0.4410*** (0.0096)
Fraction of Black Residents in Each County	0.8084*** (0.0979)
Fraction of Hispanic Residents in Each County	1.3438*** (0.1003)
Percentage of Commuters using Pub. Transportation in each county	5.0215*** (0.4140)
Log(Median Income of Each County) (in U.S. dollars)	1.3387*** (0.0579)
Percentage of Senior Residents of Each County (≥ 70 years)	1.8825*** (0.5255)
Percentage of Children Residents of Each County (18 years)	0.6614 (0.4733)
Constant	− 16.6781*** (0.6462)
R_squared	0.69
Counties	2,923

Table 2. Analysis of county fixed effects. *** p < 0.01, ** p < 0.05, * p < 0.1.

fixed effects α on several demographic variables of each county. The coefficients from this regression should be thought of as the impacts of these demographics on the transmission rate. The results from this regression are reported in Table 2. We observe that the contagion in the disease is increased with greater population density and the percentage of commuters who use public transportation. We also observe that contagion rates are higher in areas with a higher fraction of Black and Hispanic residents. Furthermore, the rate of spread is higher for seniors than for younger people, but children and non-senior adults do not seem to have statistically significantly different rates of contagion.

We next measure the out-of-sample prediction accuracy of our model using a hold-out sample of 75 days (May 24–August 6) to see how well our model forecasts new cases. We use the observed county level of daily social distancing for our out-of-sample predictions. Nationally, this reflects an approximately 50–60% return-to-normalcy, but this varies quite a bit across the country. We define the percentage return-to-normalcy as $\frac{SocialDistancingPeak_i - SocialDistancing_{i,t}}{SocialDistancingPeak_i - SocialDistancingBeforeCOVID_i}$, where $SocialDistancingPeak_i$ is the social distancing level in county i at its peak (April 5–April 11, 2020), $SocialDistancingBeforeCOVID_i$ is the observed lowest level of social distancing in February, and $SocialDistancing_{i,t}$ represents social distancing level on date t. For example, a 25% towards normalcy represents social distancing at the level of 0.25×(minimum social distancing) + 0.75×(maximum social distancing).

Figure 1 shows the US actual cumulative cases along with out-of-sample forecasts from a model with $\omega = 0.57$ and a standard model with $\omega = 1$. The black hashed line represents the actual cumulative cases in the US. The green solid line and the red dashed-line show the out-of-sample forecasts with $\omega = 0.57$ and $\omega = 1$, respectively. We readily observe that the model with $\omega = 0.57$ fits the data well while the model with $\omega = 1$ does not. Three states that had their Shelter-in-Place orders expire or stuck down early are Florida, Georgia, and Wisconsin. To further evaluate our model's accuracy in prediction, we repeat the same out-of-sample prediction comparisons for these three states in Figure 2. The figure again shows that the model with $\omega = 0.57$ has a much better fit than the model with $\omega = 1$.

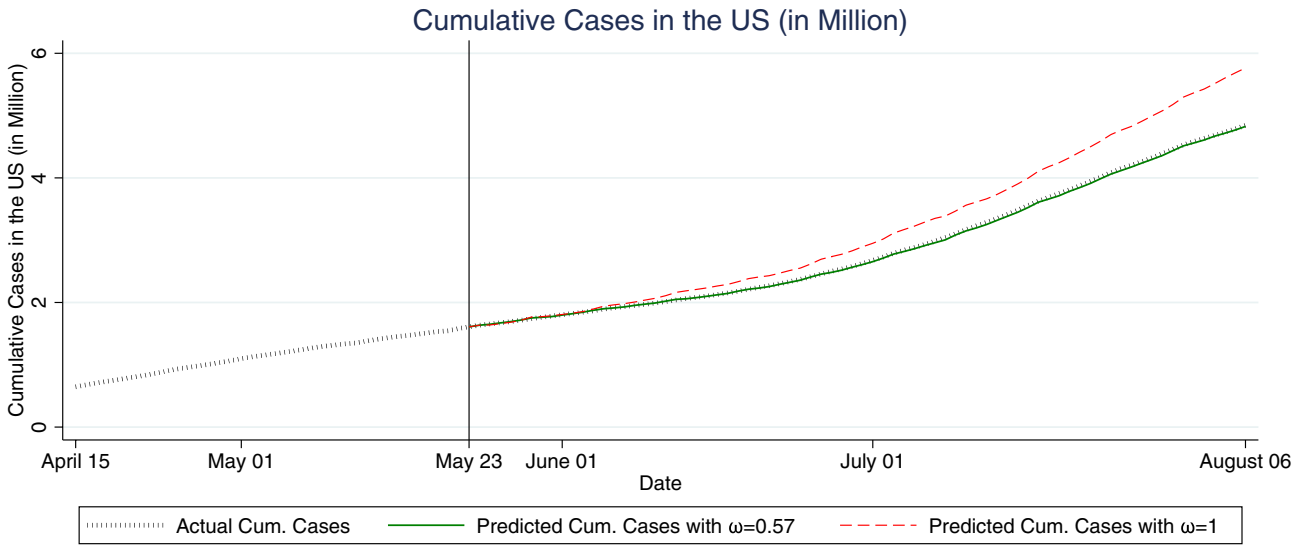

Figure 1. Out-of-sample fit comparisons of the US between our model and standard SIR model. The vertical line on May 23, 2020 indicates the last date used to estimate each model.

We next simulate daily and cumulative cases from August 7 to October 31, 2020 under different levels of social distancing. When forecasting future cases, we use previous 5-year county temperature data and the May 2020 county average humidity. The top of Figure 3 shows three sets of forecast daily cases after August 6, corresponding to 75%, current, and 25% levels of return-to-normalcy. We observe that social distancing at the current 60% return-to-normalcy first leads to a slightly increasing but then slowly decreasing number of cases, going from around 55,000 cases per day in early August to 25,000 cases per day in the end of October. If the US practiced social distancing at the level reflecting a 25% return-to-normalcy for even a few weeks, new cases would drop to a much lower level of around 9,000 per day. On the other hand, a return to a 75% level of the normalcy would cause cases to surge for about two months. The pattern of the surge is consistent with recent studies on the relaxation of non-pharmaceutical interventions such as shelter-in-place orders[18]. However, after two months cases would again reach a long-term plateau, although this would occur at a level that was almost double of what would be experienced under the early-August level of social distancing. The bottom of Figure 3 depicts the corresponding cumulative cases for the same time period under 100%, 75%, current, 25%, and 0% levels of return-to-normalcy. The figure shows a consistent pattern where the cumulative cases look almost linear after the initial take-offs. There would be substantially more cases if we returned to the pre-COVID level of social distancing.

Methods

In this subsection, we detail the assumptions we make and the estimation procedure. The model is laid out in Eqs. (1) and (2) above. For simplicity, we rewrite Eq. (2) as $R_{i,t} = \exp\left(X'_{i,t}\Phi + \varepsilon_{i,t}\right)$, where $X_{i,t}$ includes county dummy variables, date dummy variables, the measure of social distancing $d_{i,t}$, and daily average temperature $m_{i,t}$ and humidity $h_{i,t}$. Φ is the vector containing the parameters α, β, λ, μ, and θ, which measures the impact of each element in the vector $X_{i,t}$ on the transmission rate $R_{i,t}$. 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:

$$\left[\ln\left(y_{i,t}\right) - \ln\left(S_{i,t}\right)\right] = X'_{i,t}\Phi + \omega \ln\left(Y_{i,t-2} - Y_{i,t-8}\right) + \varepsilon_{i,t} \tag{3}$$

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:

$$\left[\ln\left(y_{i,t}+1\right) - \ln\left(S_{i,t}\right)\right] = X'_{i,t}\Phi + \omega \ln\left(Y_{i,t-2} - Y_{i,t-8}\right) + \varepsilon_{i,t} \tag{4}$$

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 social distancing levels (and regulations) are not determined in a vacuum: Rather, people social distance more in areas that are hit harder by COVID-19. Thus, $\varepsilon_{i,t}$ may be correlated with social distancing, causing a biased measurement of the impact of social distancing on the rate of contagion. We thus use an instrumental variables (IV) technique to control for this endogeneity bias, where the amount of rain is our instrument for social distancing. Specifically, we assume that rain directly shifts the level of social distancing, but is not correlated with $\varepsilon_{i,t}$ conditional on the temperature and humidity. Several other papers have used rain as an instrument for social distancing[19–22]. The first-stage F-statistic for the strength of rain as an instrument is 214.44, which is highly significant, indicating that rain is a strong instrument.

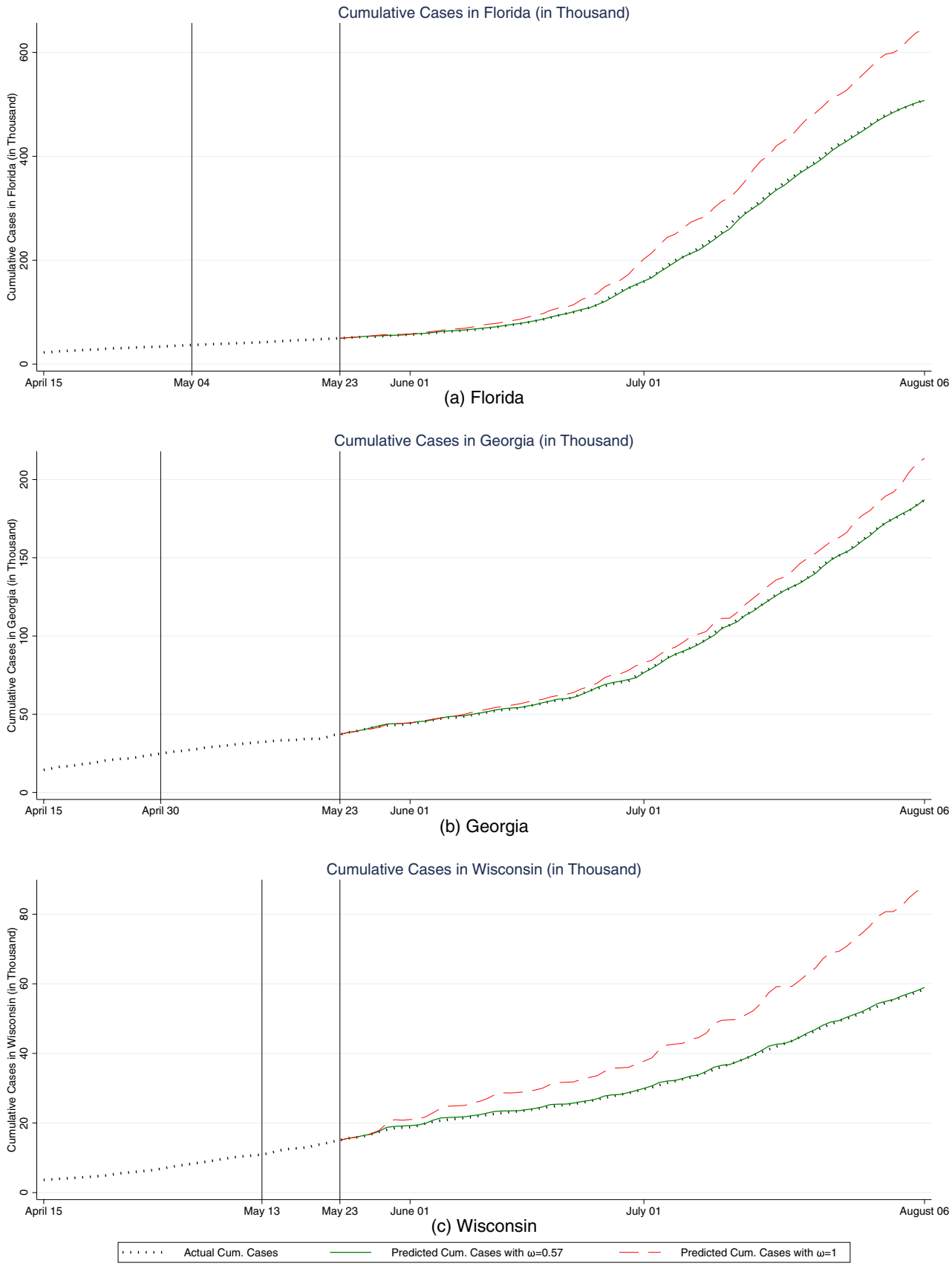

Figure 2. Out-of-sample fit comparisons of Florida, Georgia, and Wisconsin between our model and the standard SIR model. The vertical line on May 23, 2020 indicates the last date used to estimate each model. The vertical lines to the left indicate the expiration date of Shelter-in-Place order in each state.

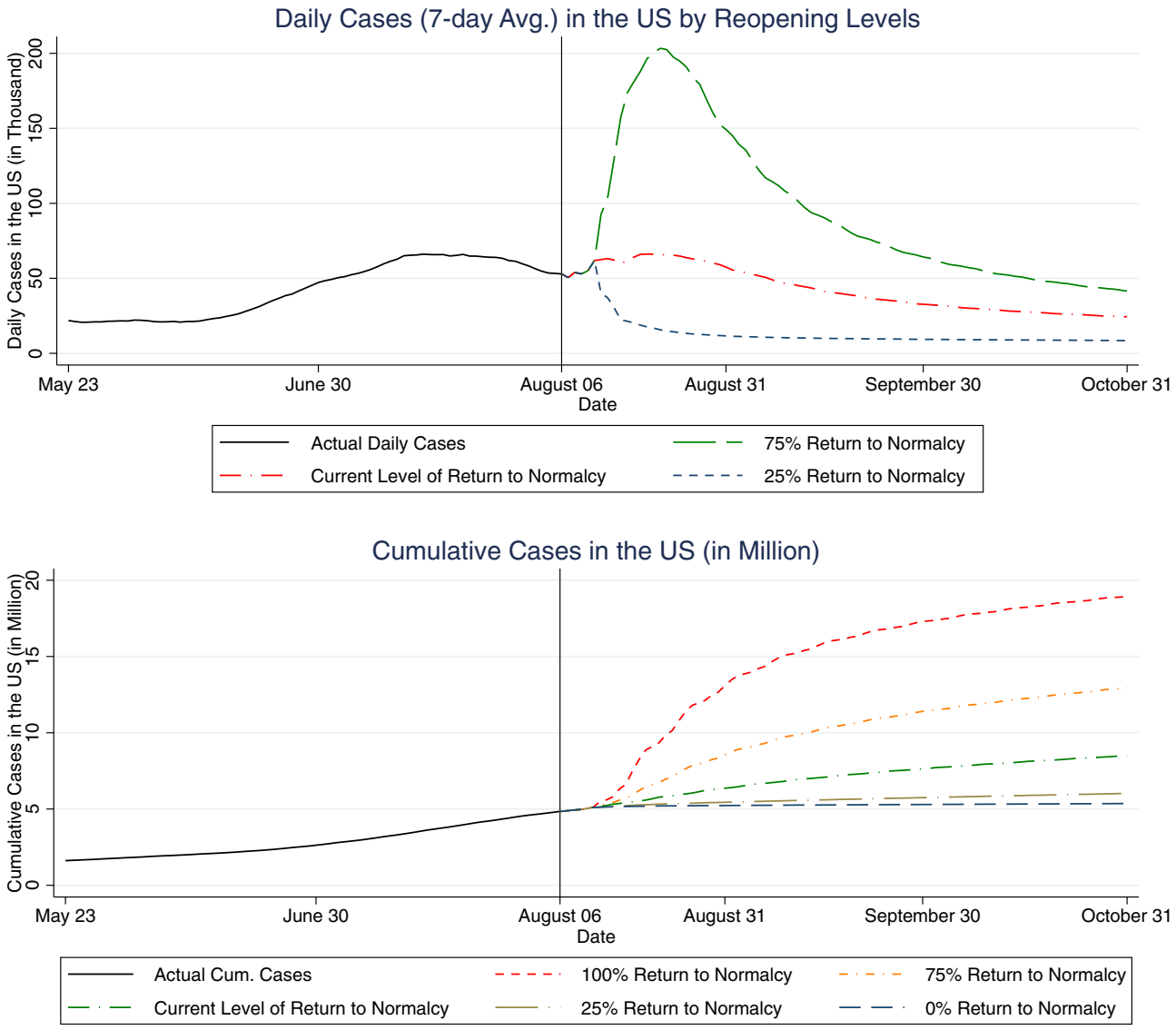

Figure 3. Daily and cumulative cases forecasting under different reopening strategies. The vertical line indicates the last date of case data sample.

Data

Our data come from a multitude of sources. We detail the data sources at https://github.com/songyao21/covid_data_depot. There are a few nonstandard issues to note. Our data on COVID-19 cases consists of county-level, officially confirmed daily case data of 2,924 US counties from February 1 to August 6 (with the last 75 days used as a hold-out sample). COVID-19 also has an incubation period of approximately 5 days[23,24]. Because of this lag from infection to diagnosis, we assume that 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%[25]. We also assume that any deaths occur 14 days after the confirmed test results. On May 23, 2020, the last day of our estimation case data, there were 92,622 deaths in the US. On May 9, 2020, there were 1,304,726 officially diagnosed cases. We hence obtain the factor as (92,622/0.0075)/1,304,726 = 9.5. We round this number up to 10. This is consistent with Centers for Disease Control and Prevention (CDC) director Robert Redfield's estimate of the ratio between actual and confirmed cases[26]. Our estimates are not sensitive to the specific factor we use. When we run the simulations, we divide our model's predicted case numbers by 10, which gives us the prediction of diagnosed cases.

Conclusion

We use a modified SIR model to study the impacts of different factors on the spread of COVID-19. We find that the impact of each additional infectious individual decreases as more people become infected. A potential mechanism underpinning this finding is that infections are more likely to occur within interconnected networks. Understanding the shape of this relationship, and the nonlinear aspects of it, are important for understanding

how COVID-19 spreads. Unlike previously-estimated SIR models, our model allows for the possibility that the contagion process will grow or shrink at relatively steady levels, whereas traditional SIR models have contagion either taking off exponentially (if $R > 1$) or falling quickly (if $R < 1$).

We further find that social distancing helps to curb the speed of the spread. Consequently, we need to be cautious of breakouts in networks and maintain a reasonably high level of social distancing during the reopening of the economy. Taking the network effects and social distancing effects together gives more accurate forecasts about the timeline of the disease spread, and the ability to analyze and set policies about when to instate shelter-in-place restrictions or when to allow businesses to be open.

Received: 11 August 2020; Accepted: 9 November 2020
Published online: 23 November 2020

References

1. World Health Organization coronavirus disease (COVID-19) dashboard. https://covid19.who.int/. Accessed on October 27, 2020.
2. The Opportunity Insights economic tracker. https://tracktherecovery.org/. Accessed on October 27, 2020.
3. Google COVID-19 community mobility reports. https://www.google.com/covid19/mobility. Accessed on October 27, 2020.
4. Wu, Z. & McGoogan, J. M. Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in China. *JAMA* **323**, 1239–1242. https://doi.org/10.1001/jama.2020.2648 (2020).
5. Chen, T. *et al.* Clinical characteristics of 113 deceased patients with coronavirus disease 2019: Retrospective study. *BMJ* **368**, 1–12. https://doi.org/10.1136/bmj.m1091 (2020).
6. Turk, C., Turk, S., Malkan, U. & Haznedaroglu, I. Three critical clinicobiological phases of the human sars-associated coronavirus infections. *Eur. Rev. Med. Pharmacol. Sci.* **24**, 8606–8620 (2020).
7. Ajuria-Illarramendi, O., Martinez-Lorca, A. & del Prado Orduña-Diez, M. [18f]fdg-pet/ct in different covid-19 phases. *IDCases* **21**, 1–2. https://doi.org/10.1016/j.idcr.2020.e00869 (2020).
8. Alsuliman, T., Alasadi, L., Alkharat, B., Srour, M. & Alrstom, A. A review of potential treatments to date in covid-19 patients according to the stage of the disease. *Curr. Res. Transl. Med.* **68**, 93–104. https://doi.org/10.1016/j.retram.2020.05.004 (2020).
9. Sun, G.-Q. *et al.* Transmission dynamics of covid-19 in Wuhan, China: effects of lockdown and medical resources. *Nonlinear Dyn.* **24**, 1–13. https://doi.org/10.1007/s11071-020-05770-9 (2020).
10. Li, M. *et al.* Analysis of covid-19 transmission in Shanxi province with discrete time imported cases. *Math. Biosci. Eng.* **17**, 3710–3720. https://doi.org/10.3934/MBE.2020208 (2020).
11. Du, Z. *et al.* Risk for transportation of coronavirus disease from Wuhan to other cities in China. *Emerging Infect. Diseases* **26**, 1049–1052. https://doi.org/10.3201/eid2605.200146 (2020).
12. Chen, T.-M. *et al.* A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infect. Diseases Poverty* **9**, 24. https://doi.org/10.1186/s40249-020-00640-3 (2020).
13. Roda, W. C., Varughese, M. B., Han, D. & Li, M. Y. Why is it difficult to accurately predict the covid-19 epidemic?. *Infect. Disease Model.* **5**, 271–281. https://doi.org/10.1016/j.idm.2020.03.001 (2020).
14. Nishiuram, H., Linton, N. M. & Akhmetzhanov, A. R. Serial interval of novel coronavirus (covid-19) infections. *Int. J. Infect. Diseases* **93**, 284–286 (2020).
15. Wang, J., Tang, K., Feng, K. & Lv, W. High temperature and high humidity reduce the transmission of covid-19. *Working Paper* (2020). Accessed on May 20, 2020 at SSRN: https://ssrn.com/abstract=3551767 or https://doi.org/10.2139/ssrn.3551767.
16. 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. *Working Paper* (2020). Accessed on May 20, 2020 at https://www.medrxiv.org/content/10.1101/2020.03.05.20031872v1.
17. Wang, M. *et al.* Temperature significant change covid-19 transmission in 429 cities. *Working Paper* (2020). https://www.medrxiv.org/content/10.1101/2020.02.22.20025791v1. Accessed on May 20, 2020.
18. Li, Y. *et al.* The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number R of SARS-CoV-2: a modelling study across 131 countries. *The Lancet Infectious Diseases* **1–10**, https://doi.org/10.1016/S1473-3099(20)30785-4 (2020).
19. Adda, J. Economic activity and the spread of viral diseases: evidence from high frequency data. *Q. J. Econ.* **131**, 891–941 (2016).
20. Qiu, Y., Chen, X. & Shi, W. Impacts of social and economic factors on the transmission of coronavirus disease 2019 (covid-19) in China. *J. Popul. Econ.* **33**, 1127–1172. https://doi.org/10.1007/s00148-020-00778-2 (2020).
21. Kapoor, R. *et al.* God is in the rain: The impact of rainfall-induced early social distancing on covid-19 outbreaks. *Available at SSRN* **3605549** (2020).
22. Holtz, D. *et al.* Interdependence and the cost of uncoordinated responses to covid-19. *Proc. Nat. Acad. Sci.* **117**, 19837–19843. https://doi.org/10.1073/pnas.2009522117 (2020).
23. Lauer, S. A. *et al.* The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Ann. Internal Med.* **172**, 577–582 (2020).
24. Li, Q. *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1200–1207 (2020).
25. Meyerowitz-Katz, G. & Merone, L. A systematic review and meta-analysis of published research data on covid-19 infection-fatality rates. *Working Paper* (2020).
26. Sun, L. H., Sun, L. H. & Achenbach, J. CDC chief says coronavirus cases may be 10 times higher than reported. *Washington Post* (2020). Accessed on June 25, 2020 at https://www.washingtonpost.com/health/2020/06/25/coronavirus-cases-10-times-larger.

Acknowledgements

We thank SafeGraph Inc. for providing the mobility data used in our analysis.

Author contributions

M.L., R.T., and S.Y. formulated the model, crafted the data analysis strategy, and drafted and revised the manuscript. M.L. and S.Y. conducted the data analysis.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-77292-8.

Correspondence and requests for materials should be addressed to S.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.