

Optimal inference of the start of COVID-19

Zheng-Meng Zhai,^{1,*} Yong-Shang Long,^{1,*} Ming Tang^{①,1,2,†} Zonghua Liu,¹ and Ying-Cheng Lai^{②,3,4,‡}¹State Key Laboratory of Precision Spectroscopy and School of Physics and Electronic Science,
East China Normal University, Shanghai 200241, China²Shanghai Key Laboratory of Multidimensional Information Processing,
East China Normal University, Shanghai 200241, China³School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287, USA⁴Department of Physics, Arizona State University, Tempe, Arizona 85287, USA

(Received 28 August 2020; accepted 28 January 2021; published 17 February 2021)

According to the official report, the first case of COVID-19 and the first death in the United States occurred on January 20 and February 29, 2020, respectively. On April 21, California reported that the first death in the state occurred on February 6, implying that community spreading of COVID-19 might have started earlier than previously thought. Exactly what is time *zero*, i.e., when did COVID-19 emerge and begin to spread in the U.S. and other countries? We develop a comprehensive predictive modeling framework to address this question. Using available data of confirmed infections to obtain the optimal values of the key parameters, we validate the model and demonstrate its predictive power. We then carry out an inverse inference analysis to determine time zero for 10 representative states in the U.S., plus New York City, United Kingdom, Italy, and Spain. The main finding is that, in both the U.S. and Europe, COVID-19 started around the new year day.

DOI: [10.1103/PhysRevResearch.3.013155](https://doi.org/10.1103/PhysRevResearch.3.013155)

I. INTRODUCTION

The first case of COVID-19 in the United States was reported on January 20, 2020 and, according to the official account, the first death on American soil occurred on February 29. Astonishingly, on April 21, California reported that the first death in the state occurred on February 6, more than three weeks earlier than previously reported. One implication is that community spreading may have already occurred in the U.S. three weeks earlier than believed. The importance of knowing precisely the starting date of community spreading cannot be overstated: this is the date based on which government mitigating actions and control measures would be imposed. In the U.S., according to the government report, the onset of an exponential increase in the infections occurred in the middle of March, leading to the belief that COVID-19 began the phase of community spreading around the same time. Based on this perception, the White House issued a nationwide social-distancing order on March 16. Statewide stay-at-home or shelter-in-place orders were given by the governors of various states at different times. The effectiveness of these government actions notwithstanding, as of August

28, there have been close to six million cases in the US and over 181 000 deaths. This devastating development means that the perceptive date of community spreading of COVID-19 in the U.S. was wrong: it could have been weeks earlier than the governments have chosen to believe.

To develop a reliable method to infer the starting date of COVID-19, or any infectious disease, is of uttermost importance. Precise knowledge of exactly when community spreading started would prompt the government to take actions at the earliest possible moment, drastically reducing the number of infections and saving many thousands of lives. More specifically, knowing this time in combination with knowledge about the symptomatic individuals in the early stage of the disease spreading enables (1) an effective reduction in the range of contact tracing, which increases the chance of accurately locating the source of infection with limited resources, (2) an assessment of the ability to infect of the virus and the way by which it spreads, providing guidance for early control measures, (3) determination of the interstate and international propagation paths of the virus, and (4) providing unequivocal early warnings for the governments. In this regard, a recent work based on gene sequencing analysis found clusters of related viruses in patients living in different neighborhoods of New York City, suggesting that multiple, independent but isolated introductions of the virus had mainly come from Europe and other parts of the U.S. [1]. In another study [2], the frequencies of the key words such as coughing and fevers on Twitter were used for model analysis, with the finding that the actual outbreak time can be 5–19 days earlier than officially reported.

Our method is based on inverse inference and enables the starting date of the epidemic to be deduced from limited

*These authors contributed equally to this work.

†tangminghan007@gmail.com

‡Ying-Cheng.Lai@asu.edu

available data. Another feature of our model is that it contains sufficient details to capture the key dynamical behaviors of COVID-19 spreading. In particular, our inverse method of inference is based on a comprehensive, non-Markovian spreading model tailored to COVID-19 in either an open or a closed setting [3]. The dynamics is described by the time evolution of the populations in five distinct states (*SHIJR*): susceptible (*S*), hidden (*H*), infected (*I*), confirmed (*J*), and removed (*R*). The *S*, *I*, and *R* states are conventional, but the *H* and *J* states are COVID-19 proper. Of particular importance is the *H*-state population: it is the population of individuals who have already contracted the coronavirus but have shown only mild symptoms or would never show any symptoms. To evolve the dynamics, the initial hidden population, $H(t_0)$, is an essential parameter, where t_0 is the time (day) at which the number of real confirmed cases begins to increase. Government control measures are usually imposed some time after this day. It is important to note that t_0 is *not* the day when the coronavirus first appears in the community (i.e., the starting date of community spreading): the latter could be significantly earlier, at a date termed as time *zero*, denoted as $\mathbf{0}$! For a time period between t_0 and t_1 , where $t_1 > t_0$, there is an appreciable and continuous increase in the number of infections, prompting the government to impose vigorous control measures on day t_1 . The time line is thus $\mathbf{0} < t_0 < t_1$. The problem is to determine time $\mathbf{0}$. Our inverse inference method is articulated to solve this problem. By generating dynamical evolution of the epidemic model, comparing the model prediction with the limited data available after t_0 , and invoking an optimization procedure, we determine the key model parameter values including $H(t_0)$. Running the model between a hypothesized time $\mathbf{0}$ and t_0 to determine how long it takes for $H(\mathbf{0})$ (a small positive integer, e.g., one or two) to reach $H(t_0)$ allows us to pin down time $\mathbf{0}$ precisely. The relevant dates and time line are illustrated in Sec. II.

We apply the inverse method to 10 states in the U.S. plus New York City and the country as a whole, and find that time zero is as early as the beginning of January. In the U.S., the day t_1 is March 16, while t_0 varies among the individual states (e.g., February 26 for Washington and March 2 for New York). For a specific system (a state or a country), letting $\Delta T \equiv t_0 - \mathbf{0}$ be the time span between the date on which the virus started community spreading and the date of the number of real confirmed cases beginning to increase, we find an exponential scaling relation between $H(t_0)$, the number of people who already carried the coronavirus on the officially confirmed date in that system and ΔT . This means that a longer delay in reporting the first case would lead to an exponentially large virus-carrying population, rendering significantly more challenging to fully control the disease spreading. The need for early, preemptive testing thus cannot be overemphasized.

Our model has the power to predict the occurrence of community spreading of COVID-19 long before the time of official report of the outbreak, making it possible for the governments to impose control measures and to summon the essential medical resources. Take the example of the U.S. In February, there was already unmistakable evidence that the coronavirus already existed in the U.S. In complete hindsight, consider the fictitious scenario that widespread tests

had been carried out in the U.S. in February so that adequate data had been available. Inverse inference could have been carried out then to determine time zero. This could have sent the vital message to the governments that community spreading of COVID-19 had already started weeks ago. If strict government control actions had been taken at the end of February, the epidemic picture of the U.S. today would have been drastically different. Our framework of modeling and inverse inference can arguably be a valuable asset for guiding the governments to take appropriate actions for possible future outbreaks of coronavirus or other infectious diseases.

II. METHODS

In the U.S., the circumstances under which COVID-19 spreads vary dramatically among different states: not only are the levels of travel restriction orders dissimilar, but other factors affecting the disease spreading such as the population, medical resources, and social and political cultures are also distinct. A quantitative assessment of the effects of the control measures taken by the government to contain COVID-19 thus needs to be carried out on a state-by-state basis. A complication is that each individual state is not a closed system: people move into and out of the state on a daily basis. This presents a tremendous challenge to modeling [4,5], as most current data analyses and models for COVID-19 were for the setting of a closed system without considering the inbound and outbound population movements [6–27]. Another feature of the COVID-19 pandemic that most existing models did not take into account is the non-Markovian nature of the spreading dynamics, as characterized by the various time delays associated with the dynamical states. To account for the non-Markovian characteristics in the model, coupled differential equations with distinct time delays are necessary [3].

To determine time $\mathbf{0}$ for a state in the U.S. treated as an open system, we develop a coupled, dual-system spreading model. In particular, we treat the target system (A) as one under influences from another, much larger system (B) that represents all the other states. Because the size of B is much larger than that of A, in terms of the spreading dynamics, system B can be regarded as a closed system. From the standpoint of nonlinear physics, the influences of system B on system A can be viewed as a perturbation or background noise, while the effects of A on B can be neglected. The perturbation can be estimated based on the population of the target state and the empirical human movement data. The backbone of this unidirectionally coupled modeling framework is a non-Markovian spreading model incorporating various time delays in a closed-system setting [3].

For the five-state (*SHIJR*) model in the closed-system setting, there are three parameters whose values are to be determined from data: (1) the infection rate β (the probability that an individual in the *S* state catches the virus and switches into the *H* state), (2) $H(t_0)$ (the hidden population at the initial time t_0 when the available data, i.e., the number of confirmed cases, began to increase with time to enable reliable estimation of the model parameters), and (3) the fraction of undocumented infections, denoted as η , whose value is determined by the testing and surveillance capability of the government. For COVID-19, the state transitions in the

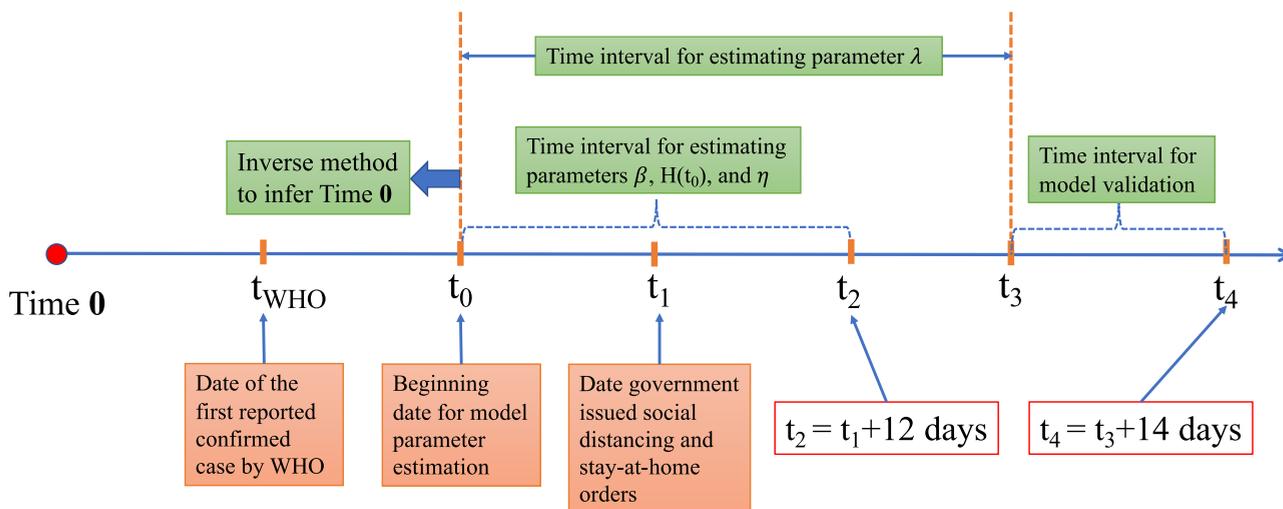


FIG. 1. Dates and timeline of inferring time $\mathbf{0}$ for a given state or country. The time t_{WHO} is the date of first emergence of case(s) reported by the World Health Organization (WHO). Time t_0 is the date when the number of confirmed cases began to increase, which is used as the initial time for optimal model parameter estimation. The time t_0 so chosen is usually about the same or later than the date of the official report of the first case(s), as there are situations where the number of cases remains unchanged for a number of days after the official date [see Table IV in Appendix B for dates of the first confirmed case(s) reported by the World Health Organization for each state/city/country studied in this paper]. Time t_1 is the date on which the government imposed control measures. The time interval $[t_0, t_2]$ is used to estimate the three key parameters of the *SHIJR* spreading model: $[\beta, H(t_0), \eta]$, where $t_2 = t_1 + 12$ days. The time interval $[t_0, t_3]$ is used for estimating λ , the parameter characterizing the effects of government control measures. The time interval $[t_3, t_4]$ is used to validate the model and demonstrate its predictive power, where $t_4 = t_3 + 14$ days.

SHIJR model are non-Markovian. In the open-system setting, an additional feature exists: there is time dependence due to human movements in and out of the system. For both closed- and open-system models, the government control measures result in an exponential decrease in the human social and movement activities. The collective effects of these measures can be described by one parameter: the exponential decay rate of the activities, denoted by λ , where a larger rate corresponds to more stringent control measures. The value of λ can be estimated based on the available epidemic data. A detailed mathematical description of the model is presented in Appendix A.

Figure 1 explains our principle to infer time $\mathbf{0}$ in terms of a number of key dates underlying COVID-19 spreading. Some days after time $\mathbf{0}$, the first or the first few cases emerged and were officially reported. However, the number of cases is typically small, e.g., one or two, and this number can remain unchanged for a number of days. Time t_0 is the date after which the number of cases begins to increase. The government imposes control measures on date $t_1 > t_0$. Since the government control is an integral part of our model, it is necessary to use data up to some date later for determining the model parameters when the effects of the control measures have been manifested in the data. We assume that this would require at least 12 days and thus set $t_2 = t_1 + 12$ days.

To estimate the four model parameters in a computationally feasible way while ensuring accuracy, we devise the following two-step procedure. We separate the four parameters into two groups: $[\beta, H(t_0), \eta]$ and λ , where the first three are associated with the intrinsic spreading process while the last is tied to the reduction in the human movements as a result of government control measures. Because of the time delay for the effects of the measures to be manifested, the value of λ does not affect

the fitting with the data in the early stage of the disease. As a result, we first use the available data in the time interval $[t_0, t_2]$ to estimate $[\beta, H(t_0), \eta]$. For each parameter, we assign an initial range of its possible values. A large number of combinations of the three parameter values are then used to evolve the model from t_0 to generate the number of confirmed cases, $J(t)$, in the time interval $[t_0, t_2]$. Using the real data, we carry out a weighted optimization procedure to determine a set of combinations of the three parameters with minimum errors. We then choose a range for λ and uniformly distribute a number of values of λ in this range. Each λ value is combined with the already determined combinations of $[\beta, H(t_0), \eta]$ to yield an equal number of combinations $[\beta, H(t_0), \eta, \lambda]$. With all the chosen λ values, this leads to a large number of combinations of the four parameters. Finally, we carry out the same optimization procedure in the time interval $[t_0, t_3]$ with $t_3 > t_2$ to determine the combination of the four parameters with the minimum error. We choose $t_3 = t_2 + 12$ days.

With the values of the optimal parameters so estimated, the model generates the time series $J(t)$ that can be compared with the data. In the time interval $[t_0, t_3]$, the agreement is generally excellent. However, this is not indicative of the predictive power of the model because the model parameters are estimated using the data in the same time interval. To test the model for prediction, we run the model in the time interval $[t_3, t_4]$, as indicated in Fig. 1. Comparison with the real data in this time interval reveals generally quite good agreement, validating the model.

The model is now ready to be used for inferring time $\mathbf{0}$. Quite straightforwardly, for a given system (a state, a city, or a country), we set $H(\mathbf{0}) = 1$, choose a number of possible candidates for time $\mathbf{0}$, and run the model from time $\mathbf{0}$ to t_0 to test which candidate date leads to the known number $H(t_0)$:

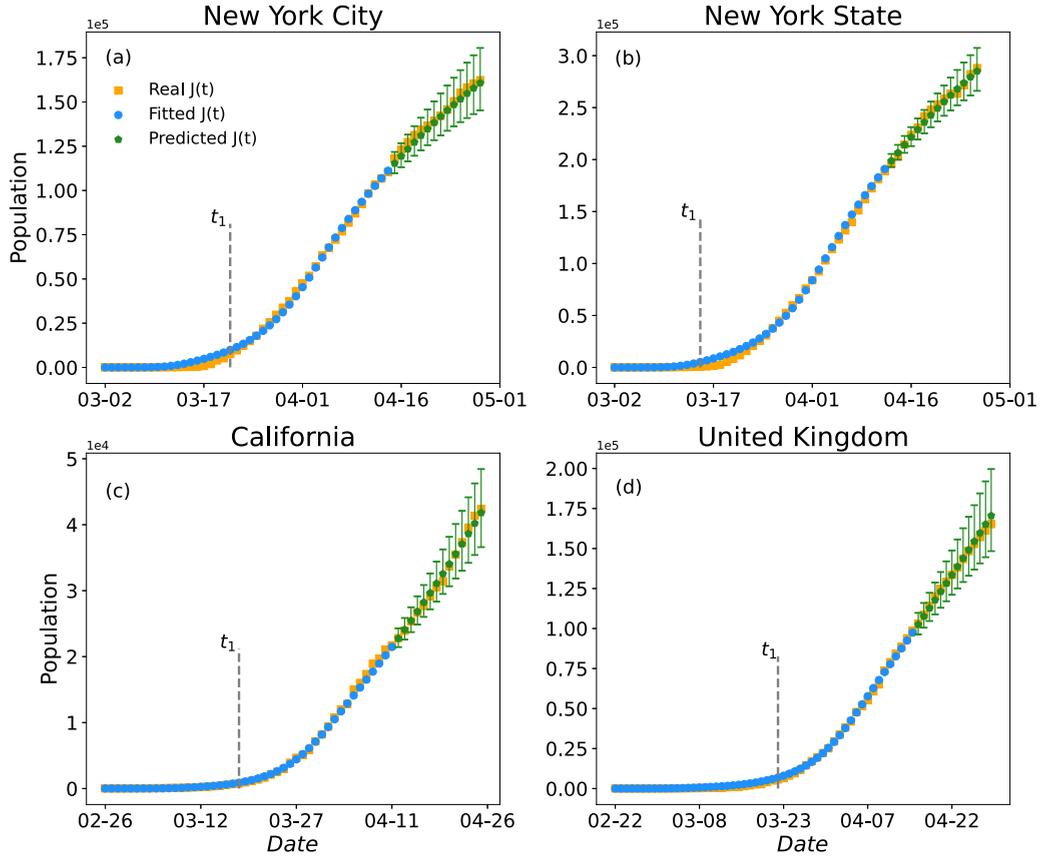


FIG. 2. Demonstration of the predictive power of the SHIJR non-Markovian model for COVID-19. Shown are the results of $J(t)$, the daily accumulative number of confirmed cases for four systems: (a) New York city treated as a closed system, (b) State of New York as an open system, (c) State of California as an open system, and (d) United Kingdom as a closed system. The orange squares are the real time series $J(t)$, the daily confirmed number of cases. The blue dots in the first phase are the model fitted $J(t)$, in which the four key model parameters are estimated based on the real data in this subtime interval (Sec. II). The green pentagons in the last 14 days are the predicted $J(t)$, whose agreement with the real data attests to the predictive power of the model. The estimated parameters are (a) $[\beta, H(t_0), \eta] = (0.19, 6000, 0.55)$ and $\lambda \in [0.098, 0.122]$; (b) $[\beta, H(t_0), \eta] = (0.19, 12000, 0.6)$ and $\lambda \in [0.111, 0.129]$ (c) $[\beta, H(t_0), \eta] = (0.22, 350, 0.65)$ and $\lambda \in [0.100, 0.120]$; (d) $[\beta, H(t_0), \eta] = (0.2, 880, 0.6)$ and $\lambda \in [0.088, 0.112]$. The range of variations in the estimated value of λ is used to generate the green error bars in the predicted $J(t)$.

the starting date that gives the correct $H(t_0)$ value is taken to be time 0.

III. RESULTS

A. Finding time zero for the U.S. states, Italy, Spain, and United Kingdom

We aim to find time zero for 10 representative states in the U.S., plus New York City (NYC), the entire U.S.A., and three European countries [Italy, Spain, and the United Kingdom (UK)]. Each State in the U.S. is treated as an open system, while NYC, U.S.A., and the three European countries are treated as a closed system. We first demonstrate that each corresponding model has the required predictive power. Figure 2 shows four representative examples in terms of the daily accumulative number of confirmed cases $J(t)$ for NYC, the state of New York, the state of California, and United Kingdom. For each example, the whole time interval is divided into two subintervals: the first (blue) is used to estimate the four key model parameters and the second subinterval (green) of 14 days is used for prediction. The model generated $J(t)$ in the

first phase is thus the result of a sophisticated, optimal fit. As can be seen from Fig. 2, our model with the parameters so estimated is capable of predicting the real data, qualifying it for inferring time zero for any given system, open or closed. Similar results for the remaining eight states in the U.S. as well as for U.S.A., Italy, and Spain are presented in Figs. 3, 4, 5.

Figure 6(a) shows the inferred time zero together with its confidence interval for the 15 states/city/countries, in an ascending order. It can be seen that SARS-CoV-2 appeared in Italy about the New Year day. In the U.S., it first emerged between January 7 and 11 in Washington or New York State. When the whole country of U.S.A. is treated as a closed system, the virus first appeared on about January 9, with confidence interval overlapping with that of the Washington and New York States, suggesting Washington or New York as the first state in which the virus emerged. It can be seen from Fig. 6(a) that the officially reported time of the emergence of the first few cases does not represent the beginning of the community spreading. In most cases, the time zeros inferred by our method were earlier than the official time, e.g., about one month earlier in Italy, two months earlier in

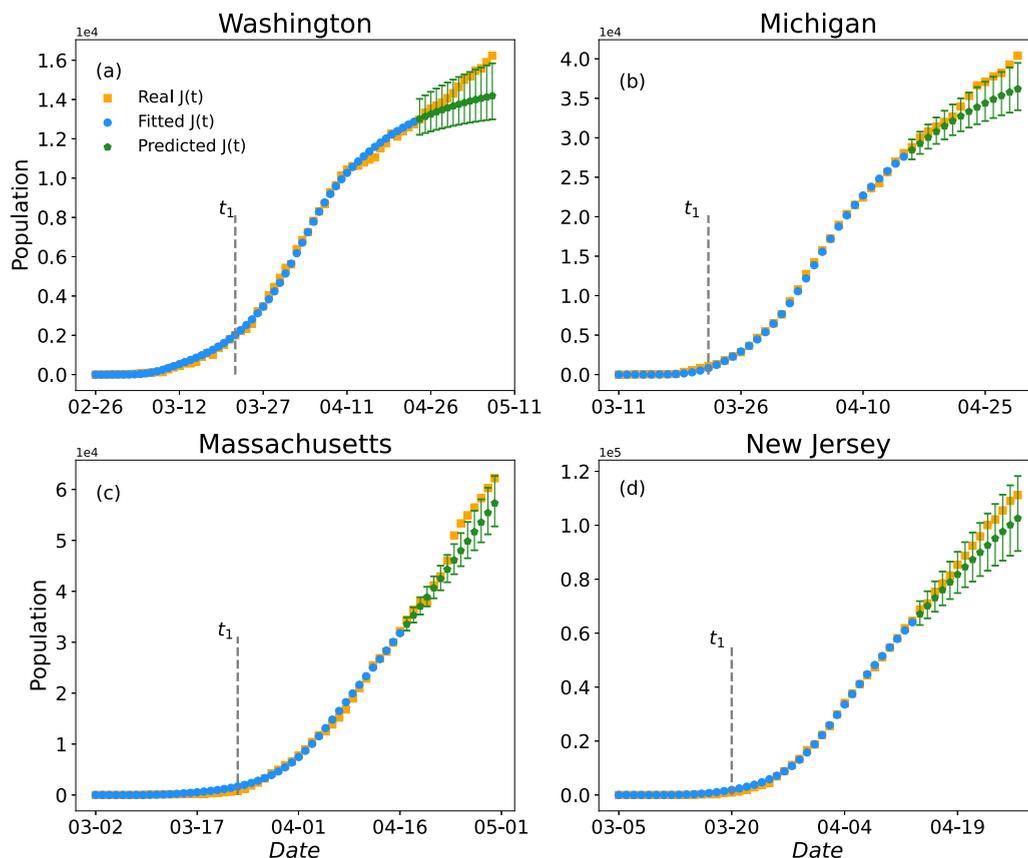


FIG. 3. Further demonstration of the model predictive power. Shown are the results of $J(t)$, the daily accumulative number of confirmed cases for four systems: (a) Washington State treated as an open system, (b) Michigan as an open system, (c) Massachusetts as an open system, and (d) New Jersey as an open system. The orange squares are the real time series $J(t)$. The blue dots in the first phase are the model fitted $J(t)$, in which the four key model parameters are estimated based on the real data in this subtime interval. The green pentagons in the last 14 days of the whole time interval are the predicted $J(t)$, whose agreement with the real data attests to the predictive power of the model. The estimated parameters are (a) $[\beta, H(t_0), \eta] = (0.16, 850, 0.6)$ and $\lambda \in [0.123, 0.157]$; (b) $[\beta, H(t_0), \eta] = (0.20, 7600, 0.8)$ and $\lambda \in [0.168, 0.193]$; (c) $[\beta, H(t_0), \eta] = (0.2, 1120, 0.75)$ and $\lambda \in [0.103, 0.117]$; (d) $[\beta, H(t_0), \eta] = (0.26, 1750, 0.55)$ and $\lambda \in [0.128, 0.153]$. The range of variations in the estimated value of λ is used to generate the green error bars in the predicted $J(t)$.

New York State and New York City, over half month earlier for Washington State. These results indicate that, before the official report of cases, COVID-19 had already spread in the community for some time. The danger of the misconception of the delayed starting date of community spread as reported by the governments is real with devastating consequences: when the governments decided to impose control measures, it may have already been too late. For example, New York State issued the lockdown order on March 22, which is more than two months later than time $\mathbf{0}$ (January 9), indicating unusually slow response of the state government to COVID-19. For the U.S. as a country, the White House issued a nationwide social-distancing order on March 16, but it was already two months later than the starting time of COVID-19 in the country, attesting to the extremely and unreasonably slow response of the federal government to the pandemic!

There are cases where the inferred dates of time zero agree approximately with the official dates, e.g., the State of Arizona. This is because the outbreak in other regions of the country (e.g., the original epicenter New York City) increased the awareness level, promoting the local governments and

population to take certain protective measures and thereby delaying the community spread.

Figure 6(b) shows that the values of the interval between time $\mathbf{0}$ and t_0 for different systems are in the range $[20, 53]$, where the maximum value of 53 days occurs for the State of New York. The average time interval $t_0 - \text{time } \mathbf{0}$ is about 33 days, indicating that SARS-CoV-2 began to spread about a month earlier than reported officially. The overall time interval of free growth from time $\mathbf{0}$ to t_1 for the States of Washington and New York as well as the whole U.S.A. is about 73 days. This explains why the controlling effect in the U.S.A. was largely unsuccessful, as the virus had spread freely for so long, making it difficult for nonpharmacological interventions such as tracing and isolation strategies to effectively inhibit the spreading. Comparing the situations between U.S.A. and Italy, the time interval $t_0 - \text{time } \mathbf{0}$ of the former is smaller than that of the latter, but U.S.A. has a greater value of the time interval $t_1 - t_0$, meaning that earlier prevention and control measures were more seriously taken in Italy. It is plausible that this resulted in better control of the virus in Italy than in the U.S.A.

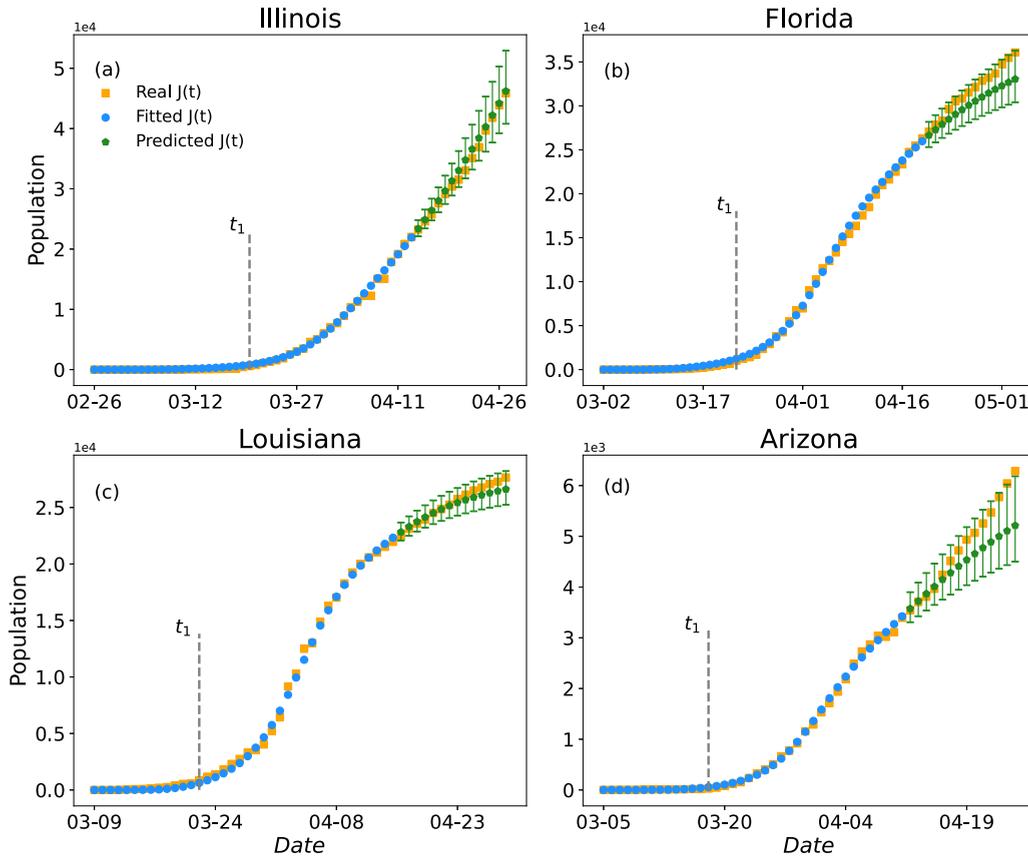


FIG. 4. Further demonstration of the model predictive power. Shown are the results of $J(t)$, the daily accumulative number of confirmed cases for four systems: (a) Illinois treated as an open system, (b) Florida as an open system, (c) Louisiana as an open system, and (d) Arizona as an open system. The estimated parameters are (a) $[\beta, H(t_0), \eta] = (0.23, 180, 0.60)$ and $\lambda \in [0.092, 0.108]$; (b) $[\beta, H(t_0), \eta] = (0.23, 440, 0.55)$ and $\lambda \in [0.150, 0.170]$; (c) $[\beta, H(t_0), \eta] = (0.25, 1850, 0.75)$ and $\lambda \in [0.239, 0.261]$; (d) $[\beta, H(t_0), \eta] = (0.28, 120, 0.7)$ and $\lambda \in [0.174, 0.206]$. Legends are the same as those of Fig. 3.

The results of time zero, together with model parameters and the estimated time line for the appearance of 5 and 20 hidden individuals for the 15 systems are summarized in Table I. Note that the values of η lie in the interval $[0.5, 0.8]$, indicating insufficient testing for a relatively long period of time after the exponential outbreak. Insufficient test, of course, gave fewer confirmed cases than actual, opening the door for the governments to undermine the severity of the pandemic and even to have the deception that COVID-19 would be under control. A consequence is that COVID-19 has continued to spread aggressively in the U.S. at the present, with no indication of control in sight. In contrast, in countries that have successfully controlled the disease, such as China and South Korea, the value of λ is between 0.1 and 0.19, signifying significantly stringent government control measures [3].

For a given system, after time zero has been determined, it is straightforward to predict how the hidden, asymptomatic population grows with time from a single case, before any government control measure is imposed. Under such a circumstance, there is free growth characterized by an exponential law, as shown in Fig. 7 on a semilogarithmic scale for the 15 systems, where the exponential growth rate is determined by the infection rate β .

What is the relation between the size of the virus-carrying hidden population at the time of first confirmed case and

ΔT , the time elapsed since zero? Figure 8 answers this question for the 15 systems. It can be seen that ΔT varies drastically among the 15 systems. When the first confirmed case was reported, there is already a sizable population of the asymptomatic individuals (hundreds or even thousands), whose actual size also varies dramatically among the systems. A general trend is that the hidden population at t_0 tends to grow exponentially with ΔT , a consequence of the free growth behavior exemplified in Fig. 7.

B. Finding time zero for the recent second COVID-19 outbreak in Beijing, China

We apply our framework of inverse inference to predict time zero for the recent second outbreak of COVID-19 in Beijing, China. From June 11 to 14, 79 cases were reported in Beijing. Since the city had had 0 new cases for several months before, the new outbreak is independent of the previous one in the January-February time frame, and can thus be regarded as a second outbreak. Analyzing the official report indicates that symptomatic patients first emerged on June 6. The extreme sparsity of the available data prevents an accurate determination of the four key parameters in our inverse model through optimization, but it is still possible to make approximate estimates. In particular, in Beijing, the regions of

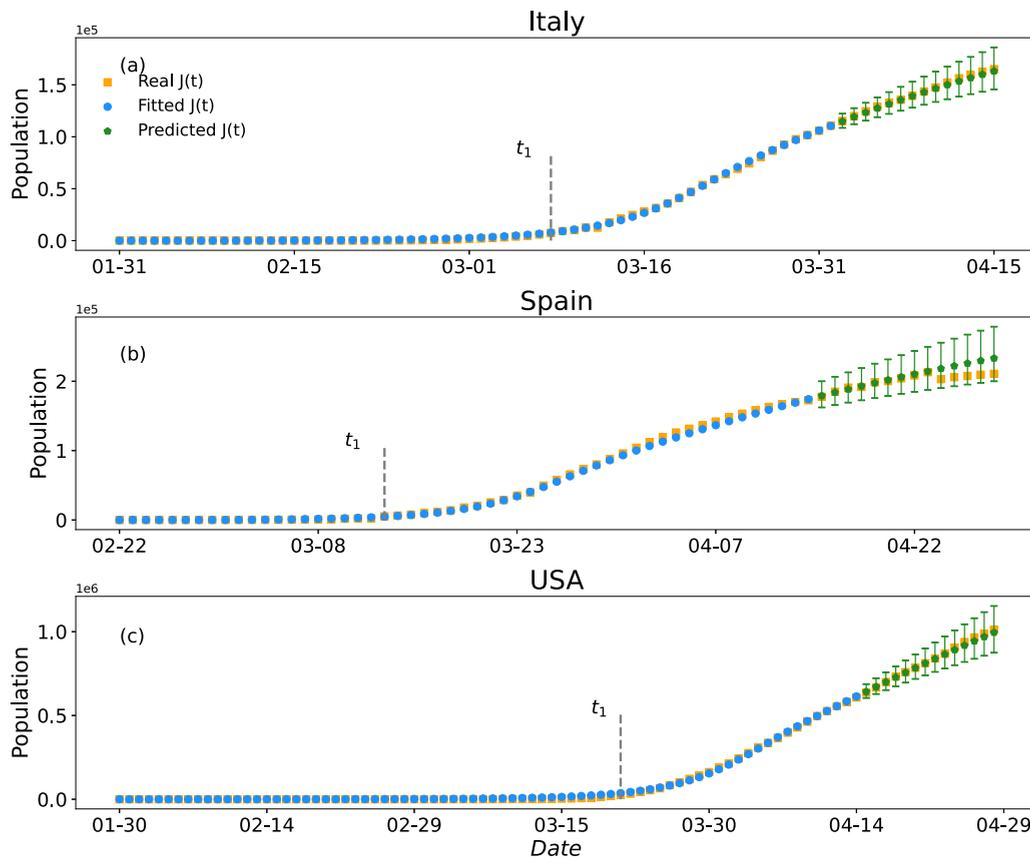


FIG. 5. Further demonstration of the model predictive power. Shown are the results of $J(t)$, the daily accumulative number of confirmed cases for four systems: (a) Italy treated as a closed system, (b) Spain as a closed system, and (c) U.S.A. as a closed system. The estimated parameters are (a) $[\beta, H(t_0), \eta] = (0.20, 330, 0.70)$ and $\lambda \in [0.127, 0.153]$; (b) $[\beta, H(t_0), \eta] = (0.25, 1400, 0.55)$ and $\lambda \in [0.138, 0.162]$; (c) $[\beta, H(t_0), \eta] = (0.22, 90, 0.55)$ and $\lambda \in [0.108, 0.132]$. Legends are the same as those of Fig. 3.

the new outbreak were isolated, and aggressive and large-scale testing was conducted immediately after the emergence of the new cases, which includes those who are asymptomatic. That

is, the individuals in the H state are tested. It is thus reasonable to hypothesize that, by June 15 (time t_0), the reported number of cases is approximately the value of $H(t)$ at this

TABLE I. Main results for 10 states in the U.S., New York City, Italy, Spain, and the United Kingdom. Time **0** is in boldface. The quantity $J(t_0)$ is the number of confirmed cases at t_0 . The parameter η is the fraction of undocumented population in the hidden state and λ measures the intensity of the government control measures. Abbreviations: It (Italy), WA (Washington State), NYC (New York City), NY (New York State), UK (United Kingdom), MI (Michigan), SP (Spain), MA (Massachusetts), CA (California), NJ (New Jersey), IL (Illinois), FL (Florida), LA (Louisiana), AZ (Arizona).

	t_0	$J(t_0)$	β	$H(t_0)$	η	λ	Time 0	$t(H = 5)$	$t(H = 20)$
It	1/31	2	0.2	330	0.7	0.14	12/28-1/4	1/7-1/12	1/11-1/15
WA	2/26	1	0.16	850	0.6	0.14	1/7-1/12	1/17-1/22	1/22-1/27
USA	1/30	5	0.22	90	0.55	0.12	1/7-1/11	1/15-1/18	1/18-1/21
NYC	3/2	1	0.19	6000	0.55	0.11	1/10-1/17	1/19-1/25	1/23-1/29
NY	3/2	1	0.19	12 000	0.6	0.12	1/6-1/12	1/15-1/20	1/18-1/24
UK	2/22	9	0.2	880	0.6	0.1	1/15-1/20	1/23-1/28	1/27-1/31
MI	3/11	2	0.2	7600	0.8	0.18	1/17-1/26	1/26-2/3	1/29-2/6
SP	2/22	2	0.25	1400	0.55	0.15	1/22-1/25	1/29-1/31	1/31-2/3
MA	3/2	1	0.2	1120	0.75	0.11	1/19-1/30	1/27-2/6	1/31-2/10
CA	2/26	10	0.22	350	0.65	0.11	1/27-2/1	2/3-2/8	2/7-2/11
NJ	3/5	2	0.26	1750	0.55	0.14	1/31-2/7	2/7-2/13	2/10-2/15
IL	2/26	2	0.23	180	0.6	0.1	2/2-2/8	2/8-2/14	2/11-2/16
FL	3/2	2	0.23	440	0.55	0.16	2/4-2/8	2/11-2/14	2/13-2/17
LA	3/9	1	0.25	1850	0.75	0.25	2/5-2/12	2/11-2/18	2/14-2/20
AZ	3/5	2	0.28	120	0.7	0.19	2/13-2/21	2/19-2/25	2/21-2/27

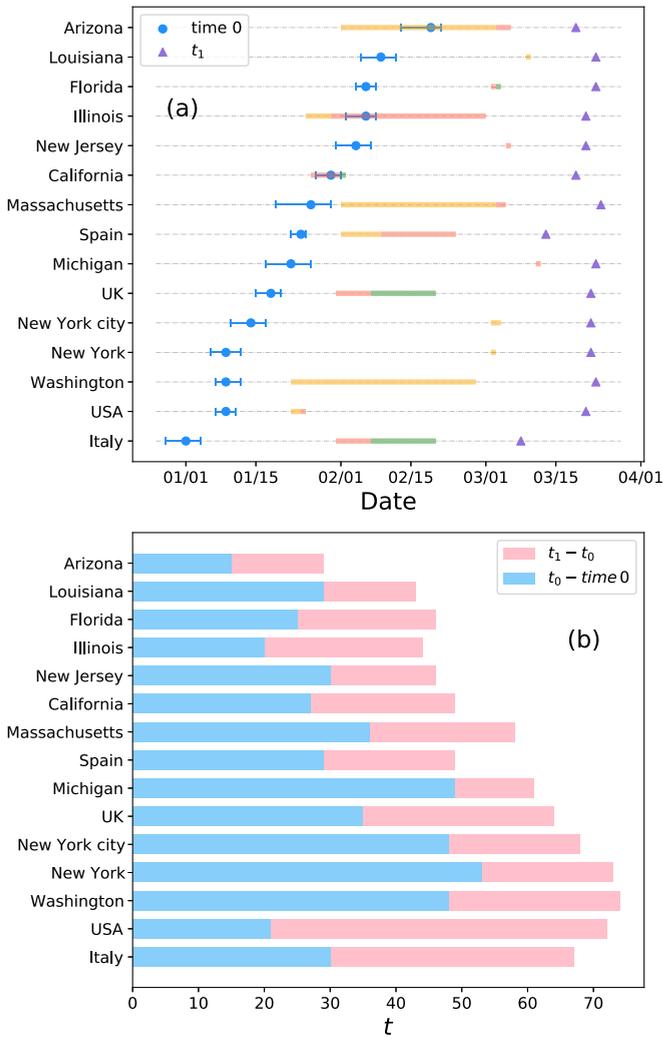


FIG. 6. Time zero for 10 states in the U.S., NYC, U.S.A. as a whole, Italy, Spain, and the United Kingdom. (a) Time 0, t_0 , and t_1 for different states/countries. The inferred dates of time 0 for the 15 systems are represented by the blue dots, together with the confidence intervals. The orange, red, and green horizontal segments represent the officially reported time duration in which there is (are) one (two and three) confirmed cases, respectively. Not every system would have all three colored segments as, e.g., more than one case could be reported by the government. Those rare cases typically represent external input, not the beginning of community spreading. The purple triangles represent the time of lockdown for the respective systems. In the U.S., the virus first emerged around January 7–11. (b) Time intervals between time 0, t_0 , and t_1 in different countries. The blue and pink bars represent the time intervals between time 0 and t_0 , and that between t_0 and t_1 , respectively.

date: $H(t_0) = 79$. Since no case was confirmed before June 15, we have $\eta = 1$. For the value of infection rate β , if we take it from the data collected in Wuhan (the epicenter in China during the first outbreak), i.e., $\beta = 0.36$, our model gives June 3 as time 0. However, if we use $\beta = 0.2$ as in the U.S. and European countries, time 0 would be May 23. For $\beta < 0.36$, time 0 would be earlier than June 3. The latest possible date of time 0 is thus June 3. Since the first symptomatic individuals appeared on June 6 and the average incubation time is about

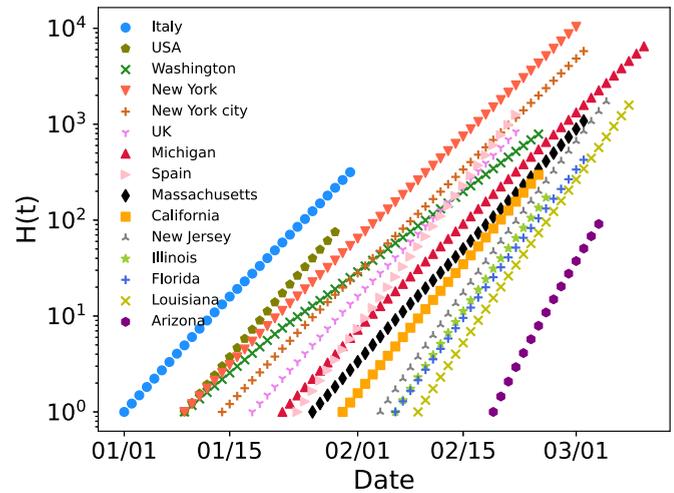


FIG. 7. Exponential growth of the hidden asymptomatic population prior to implementation of government control for the 15 systems. For uncontrolled and free growth, the exponential rate is essentially the infection rate β whose value has been estimated from data.

five days, the estimated time 0 is quite close to the actual time 0. This demonstrates that, in China where government responses are quick and testing is widely available, our model can predict time 0 even during the early stage of the outbreak with sparse data. That is, from the first official report of the outbreak, our model is already capable of providing a rough estimate of time 0, facilitating greatly localization of the spreading source(s).

IV. DISCUSSION

There were speculations that SARS-CoV-2 could have been in the U.S. a few weeks earlier than January 20, the

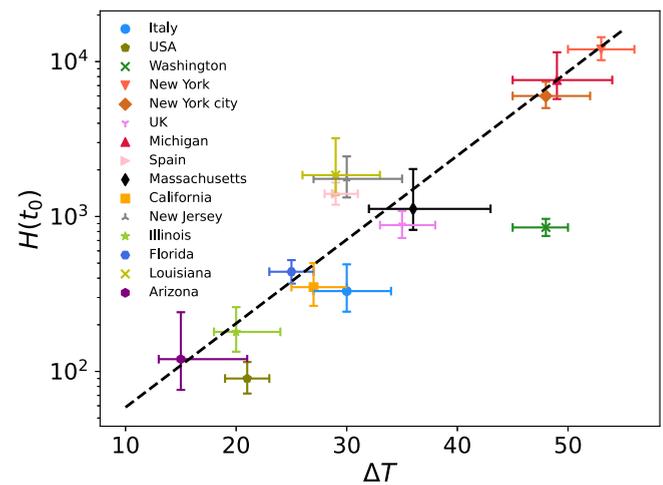


FIG. 8. Approximately exponential relation between the size of the hidden population at the time of first confirmed case reported and the time elapsed since zero. For the 15 systems studied, both the duration ΔT and the hidden population $H(t_0)$ vary widely. Among all the systems, when the first confirmed case was reported, the hidden population ranges from hundreds to thousands.

officially reported date. Inverse inference based on our non-Markovian *SHIJR* model for COVID-19 leads to a consistent answer confirming the speculations: the virus first emerged in the U.S. at the very beginning of the year. Our confidence in this result comes from our model that has been comprehensively constructed and rigorously tested in terms of the following three aspects. First, the designation of the model states and their dynamical evolution are fully in accord with the known characteristics of COVID-19, subject to government control measures. Second, the key model parameters are optimally estimated using empirical data from an adequately long time period including the initial growth phase of the epidemic. Third, the model has been validated with its predictive power firmly established through a comparison between the model generated and real data of the daily accumulative number of confirmed cases in a 14-day period that is not involved in parameter estimation. All these have been done for 10 U.S. states and New York City, the U.S. as a whole, plus three European countries most severely hit by the COVID-19 pandemic. Particularly worth noting is that our inverse procedure gives two results that are mutually consistent: the virus was already in Europe as early as the end of December 2019 and the earliest possible date for New York City to have the virus is around January 10. This consistence gives credence at a quantitative level to the widely believed proposition that the virus in New York City was from Europe through air travel [1].

Our study has demonstrated that, for a given system (a state, a city, or even a country), open or closed, our non-Markovian *SHIJR* model is capable of yielding an estimate of time zero and generating the possible epidemic trajectories into the future based on limited data with the power to predict the most likely epidemic scenario. With the inclusion of appropriate optimization and inverse inference procedures, the predictive modeling framework represents a contribution to mathematical and computational epidemiology, going beyond the existing models and offering a general and comprehensive paradigm applicable not only to COVID-19 but also to future pandemics. In addition, the framework developed in this paper can be an accurate and reliable tool or source for governments at all levels, enabling not only an accurate assessment of the government testing and surveillance capabilities for the infectious disease, but also a comprehensive evaluation of the effects of government imposed measures to control the disease. This can provide guidance for optimizing these measures to save human lives and for determining the optimal time for reopening to minimize the economical and social impacts.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grants No. 11975099, No. 11575041, No. 11675056, and No. 11835003), the Natural Science Foundation of Shanghai (Grant No. 18ZR1412200), and the Science and Technology Commission of Shanghai Municipality (Grant No. 14DZ2260800). Y.-C.L. would like to acknowledge support from the Vannevar Bush Faculty Fellowship program sponsored by the Basic Research Office of the Assistant Secretary of Defense for Research and Engineer-

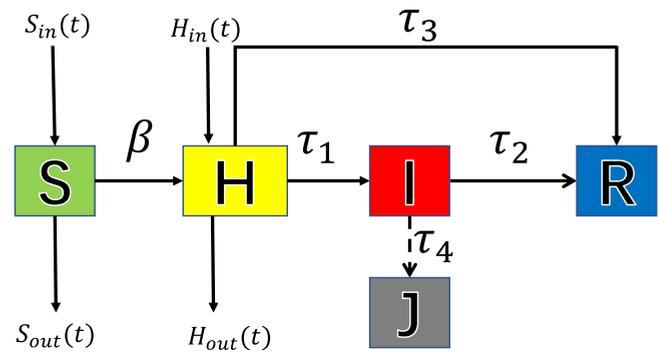


FIG. 9. Generalized COVID-19 model for an open system. Individuals in the S state are infected by H individuals at the rate β and switch to the H state. A fraction η of H -state individuals recover spontaneously or die, a process that requires τ_3 days. The parameter η thus represents the fraction of undocumented infections, and its value is determined by the testing capability of the country or state. The remaining $(1 - \eta)$ fraction of H individuals go through a transition to the I state after an average incubation period of τ_1 days, a typical non-Markovian process. With medical treatment, individuals in the I state recover or die after τ_2 days. Finally, a time delay exists for the transition from I to J : on average, the I individuals will need τ_4 days to be confirmed. The four quantities $S_{in}(t)$, $S_{out}(t)$, $H_{in}(t)$, and $H_{out}(t)$ represent the populations moving into and out of the S and H states of the open system.

ing and funded by the Office of Naval Research through Grant No. N00014-16-1-2828.

APPENDIX A: CONSTRUCTION OF NON-MARKOVIAN *SHIJR* EPIDEMIC MODEL

Figure 9 illustrates the generalized model. An individual can be in one of the five states at each time step: susceptible (S), hidden (H), infected (I), confirmed and isolated (J), and removed (R). The states S , I , and R have the same meanings as in the classical *SIR* model for infectious disease, but states H and J are unique for COVID-19. In particular, an individual in H has had the virus and is infectious but is asymptomatic or only mildly symptomatic, in contrast to the I state in which individuals show symptoms. The H individuals are capable of infecting others. The J state contains individuals who are confirmed with COVID-19. Individuals in the R state, by definition, are not infectious.

We treat each state in the U.S. as an open system, regarding the influences from all the other states as perturbations, mathematically represented by the populations moving into and out of the S and H states, denoted as $S_{in}(t)$, $S_{out}(t)$, $H_{in}(t)$, and $H_{out}(t)$, respectively, as shown in Fig. 9. These functions are determined by the travel intensity as a function of time. Two types of travel need to be distinguished: interstate and intrastate, with the corresponding intensity functions $l_{inter}(t)$ and $l_{intra}(t)$. Due to government-imposed travel restrictions, these functions decay exponentially from an initial value to a final smaller constant value. For instance, a recent estimate [17] has given that, for several major U.S. cities, the travel restrictions would reduce the outbound human movements by 50%. In general, we have $l_{inter}(t) = 1$, $e^{-\lambda_{inter}(t-t_c)}$, and $e^{-\lambda_{inter}(t_s-t_c)}$ for $t < t_c$, $t_c \leq t \leq t_s$, and $t > t_s$, respectively, and

$l_{\text{intra}}(t) = 1, e^{-\lambda_{\text{intra}}(t-t_c^i)}$, and $e^{-\lambda_{\text{intra}}(t_s^i-t_c^i)}$ for $t < t_c^i, t_c^i \leq t \leq t_s^i$, and $t > t_s^i$, respectively. For simplicity, we set $\lambda_{\text{inter}} = \lambda_{\text{intra}}$. The quantities t_c and t_s^i are, respectively, the starting dates of interstate and intrastate travel restrictions and the exponential decay in the movement activities occurs between t_c and t_s or

between t_c^i and t_s^i . The starting dates differ from state to state. The values of t_s and t_s^i are set as $t_c + 7$ and $t_c^i + 7$, respectively. Our generalized *SHIJR* model for COVID-19 epidemic for any given target state in the U.S. can be described by the following set of delayed integrodifferential equations:

$$\frac{dS(t)}{dt} = -H^n(t) - S_{\text{out}}(t) + S_{\text{in}}(t), \tag{A1}$$

$$\Phi(t) = H^n(t) - H_{\text{out}}(t) + H_{\text{in}}(t), \tag{A2}$$

$$\frac{dH(t)}{dt} = \Phi(t) - (1 - \eta) \int_{t_0}^t f_1(\tau)\Phi(t - \tau)d\tau - \eta \int_{t_0}^t f_3(\tau)\Phi(t - \tau)d\tau - (1 - \eta)f_1(t)H(t_0) - \eta f_3(t)H(t_0), \tag{A3}$$

$$\begin{aligned} \frac{dI(t)}{dt} = & (1 - \eta) \int_{t_0}^t f_1(\tau)\Phi(t - \tau)d\tau - (1 - \eta) \int_{t_0}^t f_2(\tau')d\tau' \int_{t_0}^{t-\tau'} f_1(\tau)\Phi(t - \tau' - \tau)d\tau_1 \\ & + (1 - \eta)f_1(t)H(t_0) - (1 - \eta) \int_{t_0}^t f_2(\tau)f_1(t - \tau)H(t_0)d\tau - f_2(t)I(0), \end{aligned} \tag{A4}$$

$$\begin{aligned} \frac{dR(t)}{dt} = & \eta \int_{t_0}^t f_3(\tau)\Phi(t - \tau)d\tau + (1 - \eta) \int_{t_0}^t f_2(\tau')d\tau' \int_{t_0}^{t-\tau'} f_1(\tau)\Phi(t - \tau' - \tau)d\tau_1 \\ & + \eta f_3(t)H(t_0) + (1 - \eta) \int_{t_0}^t f_2(\tau)f_1(t - \tau)H(t_0)d\tau + f_2(t)I(0) - R_{\text{out}}(t) + R_{\text{in}}(t), \end{aligned} \tag{A5}$$

$$\frac{dJ(t)}{dt} = (1 - \eta) \int_{t_0}^t f_4(\tau')d\tau' \int_{t_0}^{t-\tau'} f_1(\tau)\Phi(t - \tau' - \tau)d\tau + (1 - \eta) \int_{t_0}^t f_4(\tau)f_1(t - \tau)H(t_0)d\tau, \tag{A6}$$

$$\frac{dN}{dt} = (F_{\text{in}} - F_{\text{out}})l_{\text{inter}}(t), \tag{A7}$$

where the quantity $H^n(t)$ in the first two equations is the rate of increase in the H -state population: $H^n(t) = \beta S(t)l_{\text{intra}}H(t)/N(t)$ with $l_{\text{intra}}(t)H(t)$ representing the active H -state population that has not been isolated, $f_1(\tau)$, $f_2(\tau)$, $f_3(\tau)$, and $f_4(\tau)$ are the normal probability distribution functions of the delay time τ_1 , τ_2 , τ_3 , and τ_4 , respectively, and $N(t)$ is the population of the state as a function of time. The quantity $\Phi(t)$ is the increment of the H -state population. The input and output functions are

$$S_{\text{in}}(t) = F_{\text{in}}l_{\text{inter}}(t) \frac{S_{\text{total}}(t)}{S_{\text{total}}(t) + H_{\text{total}}(t) + R_{\text{total}}(t)},$$

$$\begin{aligned} S_{\text{out}}(t) &= F_{\text{out}}l_{\text{inter}}(t) \frac{S(t)}{S(t) + H(t) + R(t)}, \\ H_{\text{in}}(t) &= F_{\text{in}}l_{\text{inter}}(t) \frac{H_{\text{total}}(t)}{S_{\text{total}}(t) + H_{\text{total}}(t) + R_{\text{total}}(t)}, \\ H_{\text{out}}(t) &= F_{\text{out}}l_{\text{inter}}(t) \frac{H(t)}{S(t) + H(t) + R(t)}, \\ R_{\text{in}}(t) &= F_{\text{in}}l_{\text{inter}}(t) \frac{R_{\text{total}}(t)}{S_{\text{total}}(t) + H_{\text{total}}(t) + R_{\text{total}}(t)}, \\ R_{\text{out}}(t) &= F_{\text{out}}l_{\text{inter}}(t) \frac{S(t)}{S(t) + H(t) + R(t)}, \end{aligned}$$

TABLE II. The dates on which travel restriction orders were issued in the 10 states studied. For each state, the date is effectively one on which an exponential decay in intrastate traffic begins.

State	Date
Arizona	March 19, 2020
Washington	March 23, 2020
New York	March 22, 2020
New Jersey	March 21, 2020
California	March 19, 2020
Michigan	March 23, 2020
Florida	March 23, 2020
Illinois	March 21, 2020
Massachusetts	March 24, 2020
Louisiana	March 23, 2020

TABLE III. Average daily outbound and inbound populations and the total population for the 10 states studied.

State	Outbound	Inbound	Population
Arizona	13 679	6515	7 278 717
Washington	27 339	15 495	7 614 893
New York	55 012	133 095	19 453 561
New Jersey	133 507	71 056	8 882 190
California	21 171	19 979	39 512 223
Michigan	21 179	11 625	9 986 857
Florida	28 871	22 111	21 477 737
Illinois	48 160	45 451	12 671 821
Massachusetts	30 217	44 987	6 892 503
Louisiana	11 754	14 994	4 648 794

TABLE IV. Additional information and results for 10 states in the U.S., New York City, Italy, Spain, and the United Kingdom. The quantity t_{WHO} is the range of dates of first emergence of case(s) reported by the World Health Organization (WHO). Columns 3–6 list the confidence intervals of the four estimated parameters required of the non-Markovian *SHIJR* spreading model under government imposed control measures: β , $H(t_0)$, η , and λ . Abbreviations: It (Italy), WA (Washington State), NYC (New York City), NY (New York State), UK (United Kingdom), MI (Michigan), SP (Spain), MA (Massachusetts), CA (California), NJ (New Jersey), IL (Illinois), FL (Florida), LA (Louisiana), AZ (Arizona).

System	t_{WHO}	$(\beta_{\text{low}}, \beta_{\text{upp}})$	$[H_{\text{low}}(t_0), H_{\text{upp}}(t_0)]$	$(\eta_{\text{low}}, \eta_{\text{upp}})$	$(\lambda_{\text{low}}, \lambda_{\text{upp}})$
It	1/31-2/20	(0.191,0.208)	(243,491)	(0.625,0.781)	(0.127,0.153)
USA	1/22-1/25	(0.213,0.227)	(72,115)	(0.497,0.607)	0.108,0.132)
WA	1/22-2/28	(0.156,0.164)	(749,966)	(0.560,0.639)	(0.123,0.157)
NY	3/2-3/3	(0.184,0.196)	(10220,14382)	(0.552,0.653)	(0.111,0.129)
NYC	3/2-3/4	(0.182,0.198)	(5015,7405)	(0.488,0.617)	(0.098,0.122)
UK	1/31-2/7	(0.193,0.207)	(727,1084)	(0.544,0.655)	(0.088,0.112)
MI	3/11-3/12	(0.194,0.206)	(5734,11486)	(0.742,0.865)	(0.168,0.193)
SP	2/1-2/24	(0.243,0.257)	(1192,1657)	(0.514,0.587)	(0.138,0.162)
MA	2/1-3/5	(0.190,0.207)	(820,2026)	(0.678,0.852)	(0.103,0.117)
CA	1/26-2/2	(0.210,0.229)	(265,500)	(0.574,0.735)	(0.100,0.120)
NJ	3/5-3/6	(0.248,0.272)	(1327,2443)	(0.472,0.643)	(0.128,0.153)
IL	1/25-3/1	(0.220,0.240)	(134,260)	(0.518,0.694)	(0.092,0.108)
FL	3/2-3/4	(0.224,0.237)	(369,523)	(0.503,0.597)	(0.150,0.170)
LA	3/9-3/10	(0.241,0.258)	(1381,3204)	(0.680,0.844)	(0.239,0.261)
AZ	2/1-3/6	(0.263,0.296)	(76,241)	(0.590,0.826)	(0.174,0.206)

where the quantities $S_{\text{total}}(t)$, $H_{\text{total}}(t)$, and $R_{\text{total}}(t)$ are the total S -, H -, and R -state populations of the U.S. excluding the target state, $F_{\text{in}}(t)$ and $F_{\text{out}}(t)$ are the fluxes into and out of the state, which can be extrapolated from empirical data. To numerically solve the whole set of equations, the values of the initial H -state population $H(t_0)$ and of the infection rate β are needed, which can be estimated with a mathematical optimization procedure. In particular, we look for that optimal parameter combination $\hat{\Theta} = [\beta, H(t_0), \eta, \lambda]$ that minimizes the weighted difference squared between the data points $J(t_i)$ and the predicted values of $f(t_i, \Theta)$ ($i = 0, 1, \dots, n-1$), according to

$$\hat{\Theta} = \operatorname{argmin} \sum_{i=1}^n w_{t_i} [f(t_i, \Theta) - J(t_i)]^2. \quad (\text{A8})$$

Because of the necessity of assigning larger weights for more recent data, we set the weights to be $w_{t_{n-i}} = \alpha(1-\alpha)^{i-1}$ with $\alpha = 0.1$. The optimization problem can be solved to yield the optimal values β^* , $H^*(t_0)$, η^* , and λ^* using, e.g., the Levenberg-Marquardt (LM) method [28–31].

To determine the daily fluxes $F_{\text{in}}(t)$ and $F_{\text{out}}(t)$, we will use the commuting data from the U.S. Census Bureau [32], which were obtained from sampling the home and work addresses of the working population in the five-year period (2011–2015). Our estimation method is as follows. Assume that the commuting population is distributed uniformly among the states. On average, each working individual commutes 0.11 times per day. Multiplying this number by the population of the state gives the daily average number of people who commute. Denoting this number by Q and letting the commuting populations from the Census Bureau’s database be $P_{\text{in/out}}$, we obtain the ratio $C_{\text{in/out}} = Q/P_{\text{in/out}}$. Let $D_{\text{in/out}}$ be populations in and out of the state from the database. The daily fluxes can be obtained as $F_{\text{in/out}} = C_{\text{in/out}} D_{\text{in/out}}$.

For the 10 states studied, the dates t_c^i are listed in Table II and the flux values are listed in Table III.

APPENDIX B: ADDITIONAL RESULTS

Additional information and results for 10 states in the U.S., New York City, Italy, Spain, and the United Kingdom are listed in Table IV.

[1] A. S. Gonzalez-Reiche, M. M. Hernandez, M. J. Sullivan, B. Ciferri, H. Alshammery, A. Obla, S. Fabre, G. Kleiner, J. Polanco, Z. Khan, B. Albuquerque, A. van de Guchte, J. Dutta, N. Francoeur, B. S. Melo, I. Oussenko, G. Deikus, J. Soto, S. H. Sridhar, Y.-C. Wang *et al.*, Introductions and early spread of SARS-CoV-2 in the New York City area, *Science* **369**, 297 (2020).

[2] E. Gharavi, N. Nazemi, and F. Dadgostari, Early outbreak detection for proactive crisis management using Twitter data: COVID-19 a case study in the US, [arXiv:2005.00475](https://arxiv.org/abs/2005.00475).

[3] Y.-S. Long, Z.-M. Zhai, L.-L. Han, J. Kang, Y.-L. Li, Z.-H. Lin, L. Zeng, D.-Y. Wu, C.-Q. Hao, M. Tang, Z. Liu, and Y.-C. Lai, Quantitative assessment of the role of undocumented infection in the 2019 novel coronavirus (COVID-19) pandemic, [arXiv:2003.12028](https://arxiv.org/abs/2003.12028).

[4] J. R. Fauver, M. E. Petrone, E. B. Hodcroft, K. Shioda, H. Y. Ehrlich, A. G. Watts, C. B. Vogels, A. F. Brito, T. Alpert, A. Muyombwe, J. Razeq, R. Downing, N. R. Cheemarla, A. L. Wyllie, C. C. Kalinich, I. M. Ott, J. Quick, N. J. Loman, K. M. Neugebauer, A. L. Greninger *et al.*, Coast-to-coast spread of

- SARS-CoV-2 during the early epidemic in the United States, *Cell* **181**, 990 (2020).
- [5] M. U. G. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. du Plessis, N. R. Faria, R. Li, W. P. Hanage, J. S. Brownstein, M. Layan, A. Vespignani, H. Tian, C. Dye, O. G. Pybus, and S. V. Scarpino, The effect of human mobility and control measures on the COVID-19 epidemic in China, *Science* **368**, 493 (2020).
- [6] J. T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P. M. d. Salazar, B. J. Cowling, M. Lipsitch, and G. M. Leung, Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China, *Nat. Med.* **26**, 506 (2020).
- [7] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, X. Xing, N. Xiang *et al.*, Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia, *New Engl. J. Med.* **382**, 1199 (2020).
- [8] W.-J. Guan, Z.-Y. Ni, Y. Hu, W.-H. Liang, C.-Q. Ou, J.-X. He, L. Liu, H. Shan, C.-L. Lei, D. S. C. Hui *et al.*, Clinical characteristics of coronavirus disease 2019 in China, *New Engl. J. Med.* **382**, 1708 (2020).
- [9] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, A novel coronavirus outbreak of global health concern, *Lancet* **395**, 470 (2020).
- [10] J. T. Wu, K. Leung, and G. M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modeling study, *Lancet* **395**, 689 (2020).
- [11] T. Zhou, Q. Liu, Z. Yang, J. Liao, K. Yang, W. Bai, X. Lu, and W. Zhang, Preliminary prediction of the basic reproduction number of the Wuhan novel coronavirus 2019-nCoV, *J. Evid. Based Med.* **13**, 3 (2020).
- [12] Z. Du, X. Xu, Y. Wu, L. Wang, B. J. Cowling, and L. A. Meyers, The serial interval of COVID-19 from publicly reported confirmed cases, *Emerging Infect. Dis.* **26**, 1341 (2020).
- [13] N. M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, A. Dighe, I. Dorigatti, H. Fu, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, L. C. Okell, S. van Elsland, H. Thompson *et al.*, Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand (2020), report of Imperial College COVID-19 Response Team, March 16, 2020 (unpublished), <http://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/>.
- [14] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, H. Coupland, T. A. Mellan, H. Zhu, T. Berah, J. Eaton, P. N. P. Guzman, N. Schmit, L. Cilloni, K. E. C. Ainslie, M. Baguelin, I. Blake, A. Boonyasiri, O. Boyd, L. Cattarino, C. Ciavarella, L. Cooper *et al.*, Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe, *Nature* **584**, 257 (2020).
- [15] A. Aleta and Y. Moreno, Evaluation of the incidence of COVID-19 and of the efficacy of contention measures in Spain: a data-driven approach, *BMC Medicine* **18**, 157 (2020).
- [16] J. F. Rabajante, Insights from early mathematical models of 2019-nCoV acute respiratory disease (COVID-19) dynamics, *J. Environ. Sci. Manag.* **23**, 1 (2020).
- [17] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. Pastore y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini, and A. Vespignani, The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak, *Science* **368**, 395 (2020).
- [18] B. Klein, T. LaRock, S. McCabe, L. Torres, F. Privitera, B. Lake, M. U. G. Kraemer, J. S. Brownstein, D. Lazer, T. Eliassi-Rad, S. V. Scarpino, M. Chinazzi, and A. Vespignani, Assessing changes in commuting and individual mobility in major metropolitan areas in the United States during the COVID-19 outbreak, Report, Network Science Institute, Northeastern University, 2020 (unpublished), <https://www.networkscienceinstitute.org/publications/assessing-changes-in-commuting-and-individual-mobility-in-major-metropolitan-areas-in-the-united-states-during-the-covid-19-outbreak>.
- [19] A. Rădulescu, C. Williams, and K. Cavanagh, Management strategies in a SEIR-type model of COVID 19 community spread, *Sci. Rep.* **10**, 21256 (2020).
- [20] E. Dong, H. Du, and L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, *Lancet Infect. Dis.* **20**, 533 (2020).
- [21] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2), *Science* **368**, 489 (2020).
- [22] P. Liu, P. Beeler, and R. K. Chakrabarty, COVID-19 progression timeline and effectiveness of response-to-spread interventions across the United States, [medRxiv:10.1101/2020.03.17.20037770](https://doi.org/10.1101/2020.03.17.20037770).
- [23] A. J. Stier, M. G. Berman, and L. Bettencourt, COVID-19 attack rate increases with city size, [arXiv:2003.10376](https://arxiv.org/abs/2003.10376).
- [24] M. Lu, Dynamic modeling COVID-19 for comparing containment strategies in a pandemic scenario, [arXiv:2003.13997](https://arxiv.org/abs/2003.13997).
- [25] B. F. Maier and D. Brockmann, Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China, *Science* **368**, 742 (2020).
- [26] H. Tian, Y. Liu, Y. Li, C.-H. Wu, B. Chen, M. U. G. Kraemer, B. Li, J. Cai, B. Xu, Q. Yang, B. Wang, P. Yang, Y. Cui, Y. Song, P. Zheng, Q. Wang, O. N. Bjornstad, R. Yang, B. T. Grenfell, O. G. Pybus, and C. Dye, An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China, *Science* **368**, 638 (2020).
- [27] S. Hsiang, D. Allen, S. Annan-Phan, K. Bell, I. Bolliger, T. Chong, H. Druckenmiller, L. Y. Huang, A. Hultgren, E. Krasovich, P. Lau, J. Lee, E. Rolf, J. Tseng, and T. Wu, The effect of large-scale anti-contagion policies on the COVID-19 pandemic, *Nature (London)* **584**, 262 (2020).
- [28] K. Levenberg, A method for the solution of certain non-linear problems in least squares, *Q. Appl. Math.* **2**, 164 (1944).
- [29] D. W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, *J. Soc. Ind. Appl. Math.* **11**, 431 (1963).
- [30] G. Chowell, Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts, *Infect. Dis. Model.* **2**, 379 (2017).
- [31] B. Kaltenbacher, A. Neubauer, and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, (Walter de Gruyter, Berlin, 2008), Vol. 6.
- [32] <https://www.census.gov/data/tables/2015/demo/metro-micro/commuting-flows-2015.html>.