When did COVID-19 start? - Optimal inference of time ZERO

Zheng-Meng Zhai,[1,*] Yong-Shang Long,[1,*] Ming Tang,[1,2,†] Zonghua Liu,[1] and Ying-Cheng Lai[3,4,‡]

[1]*State Key Laboratory of Precision Spectroscopy and School of Physics and Electronic Science, East China Normal University, Shanghai 200241, China*

[2]*Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China*

[3]*School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287, USA*

[4]*Department of Physics, Arizona State University, Tempe, Arizona 85287, USA*

(Dated: July 24, 2020)

Abstract

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 US 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 ten representative States in the US, plus New York city, UK, Italy, and Spain. The main finding is that, in both the US and Europe, COVID-19 started around the new year day.

* These authors contributed equally to this work.

† tangminghan007@gmail.com

‡ Ying-Cheng.Lai@asu.edu

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 US 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 US, 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 time. The effectiveness of these government actions notwithstanding, as of July 24, there have been over four million cases in the US with close to 145,000 deaths. This devastating development means that the perceptive date of community spreading of COVID-19 in the US 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 infectability 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 US [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 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 are 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 **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 timeline is thus: $\mathbf{0} < t_0 < t_1$. The problem is to determine time **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 **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 **0** precisely. The relevant dates and timeline are illustrated in **Methods**.

We apply the inverse method to ten States in the US 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 US, 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, pre-emptive testing thus cannot be over-emphasized.

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 US. In February, there was already unmistakable evidence that the coronavirus already existed in the US. In complete hindsight, consider the fictitious scenario that widespread tests had been carried out in the US 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 US 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 infections diseases.

RESULTS

Finding time ZERO for the US States, Italy, Spain, and UK. We aim to find time ZERO for ten representative States in the US, plus New York City (NYC), the entire USA, and three European countries [Italy, Spain, and the United Kingdom (UK)]. Each State in the US is treated as an open system, while NYC, USA, and the three European countries are treated as a closed system (Supplementary Note of SI). We first demonstrate that each corresponding model has the required predictive power. Figure 1 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 UK. For each example, the whole time interval is divided into two sub-intervals: the first (blue) is

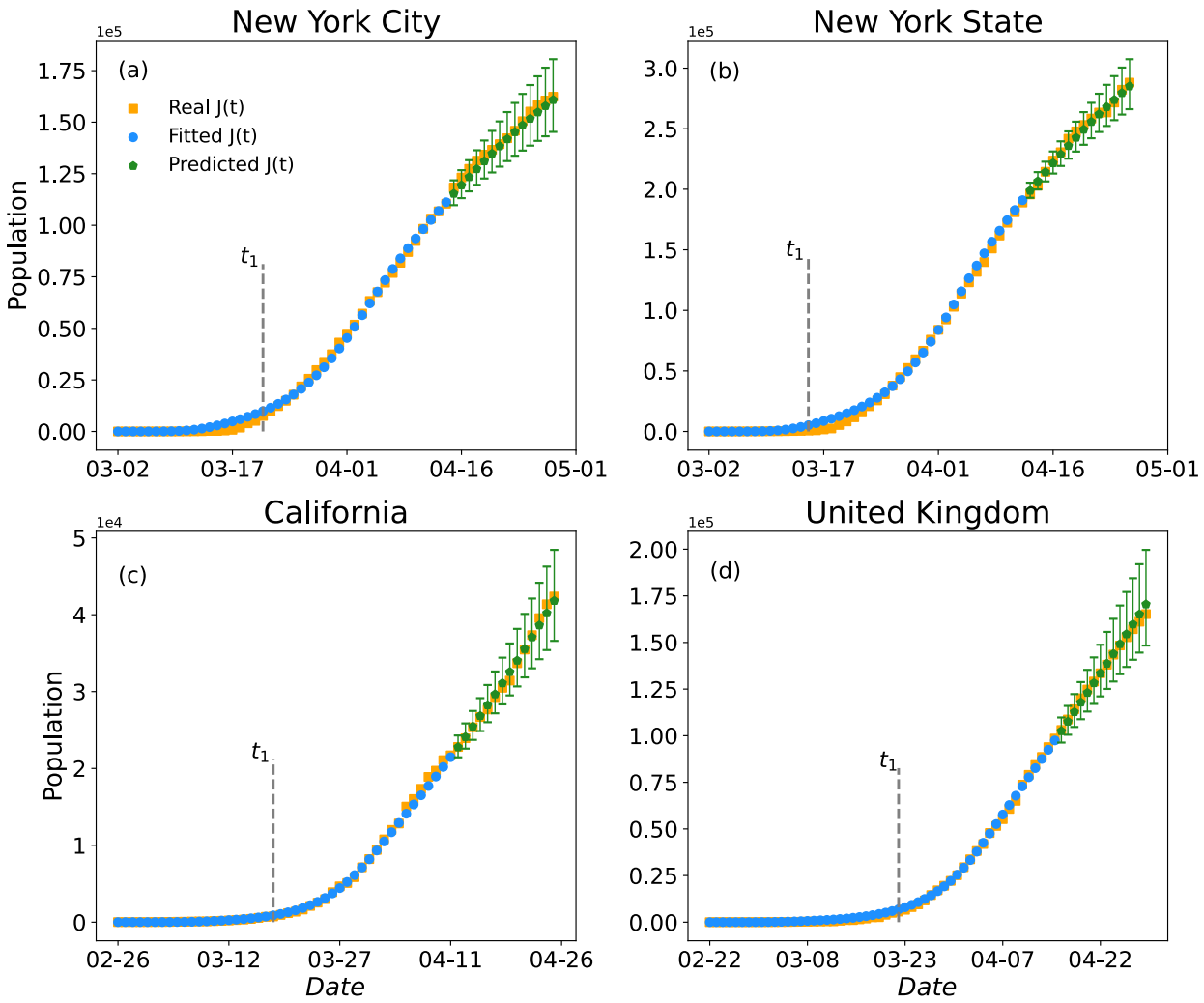

FIG. 1. Demonstration of the predictive power of the SHIJR non-Markovian model for COVID-19. Shown are the results of $J(t)$, the daily accumulative number of confirmed cases for four systems: (a) New York city treated as a closed system, (b) State of New York as an open system, (c) State of California 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 sub-time interval (**Methods**). 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 $J(t)$.

used to estimate the four key model parameters and the second sub-interval (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. 1, 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 US as well as for USA, Italy and Spain

are presented as Supplementary Figs. S1-S3.

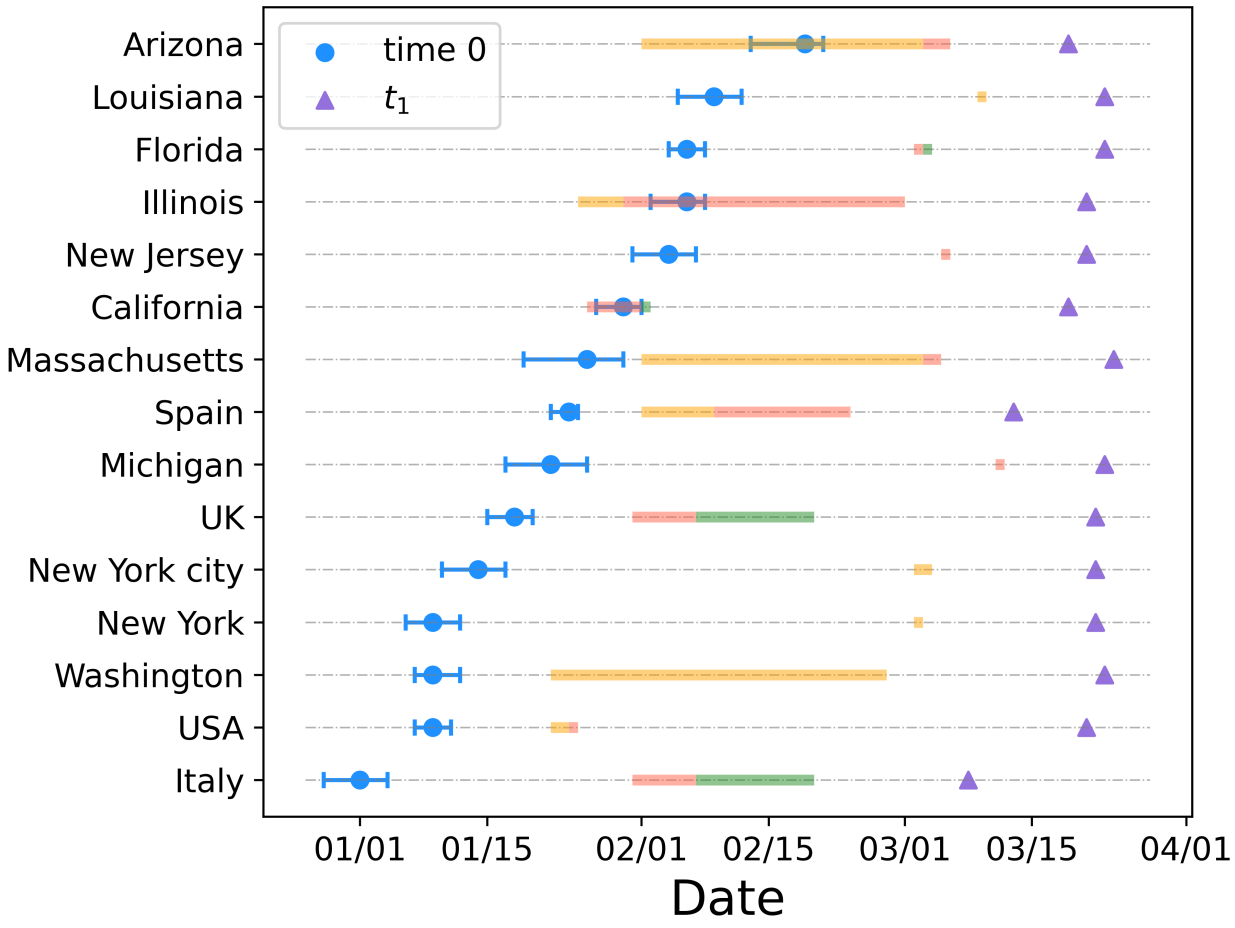

FIG. 2. Time ZERO for ten States the US, NYC, USA as a whole, Italy, Spain, and the UK. The inferred dates of time **0** for the 15 systems are represented by the blue dots, together with the confidence intervals. The orange, red, and green horizontal segments represent the officially reported time duration 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 lock-down for the respective systems. In the US, the virus first emerged around January 4-9.

Figure 2 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 the novel coronavirus appeared in Italy about the New Year day. In the US, it first emerged between January 7 and 11 in Washington or New York State. When the whole country of USA 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. 2 that the officially reported time of the emergence of the first few cases does not represent the beginning 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 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

TABLE I. Main results for ten States in the US, 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	900	0.55	0.12	**1/4-1/9**	1/14-1/19	1/20-1/24
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	10000	0.55	0.13	**1/13-1/18**	1/22-1/26	1/25-1/29
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	860	0.65	0.11	**1/25-2/2**	2/3-2/9	2/6-2/12
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	0.28	1700	0.6	0.13	**1/29-2/8**	2/4-2/13	2/7-2/15
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

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 **0** (January 9), indicating unusually slow response of the State government to COVID-19. For the US 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.

The results of time ZERO, together with model parameters and the estimated timeline for the appearance of five and 20 hidden individuals for the 15 systems are summarized in Tab. 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 US 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].

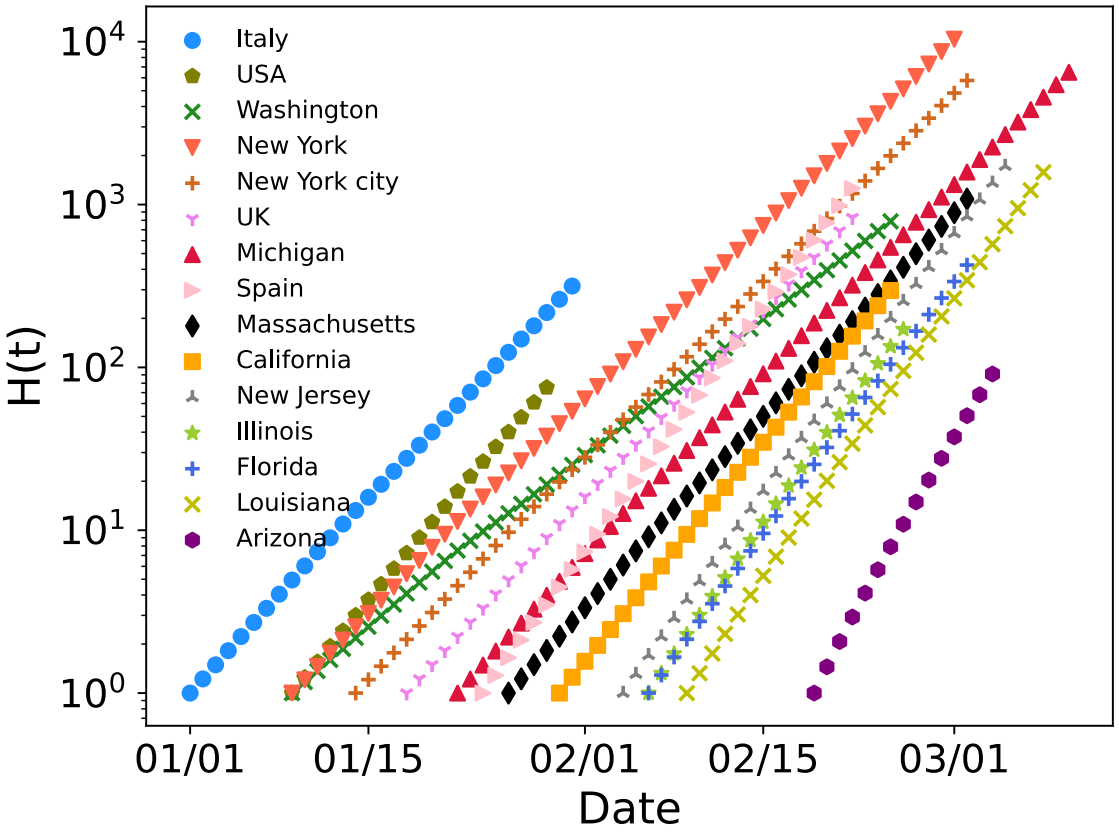

FIG. 3. Exponential growth of the hidden asymptomatic population prior to implementation of government control for the 15 systems. For uncontrolled and free growth, the exponential rate is essentially the infection rate β whose value has been estimated from data.

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. 3 on a semi-logarithmic 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 4 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. 3.

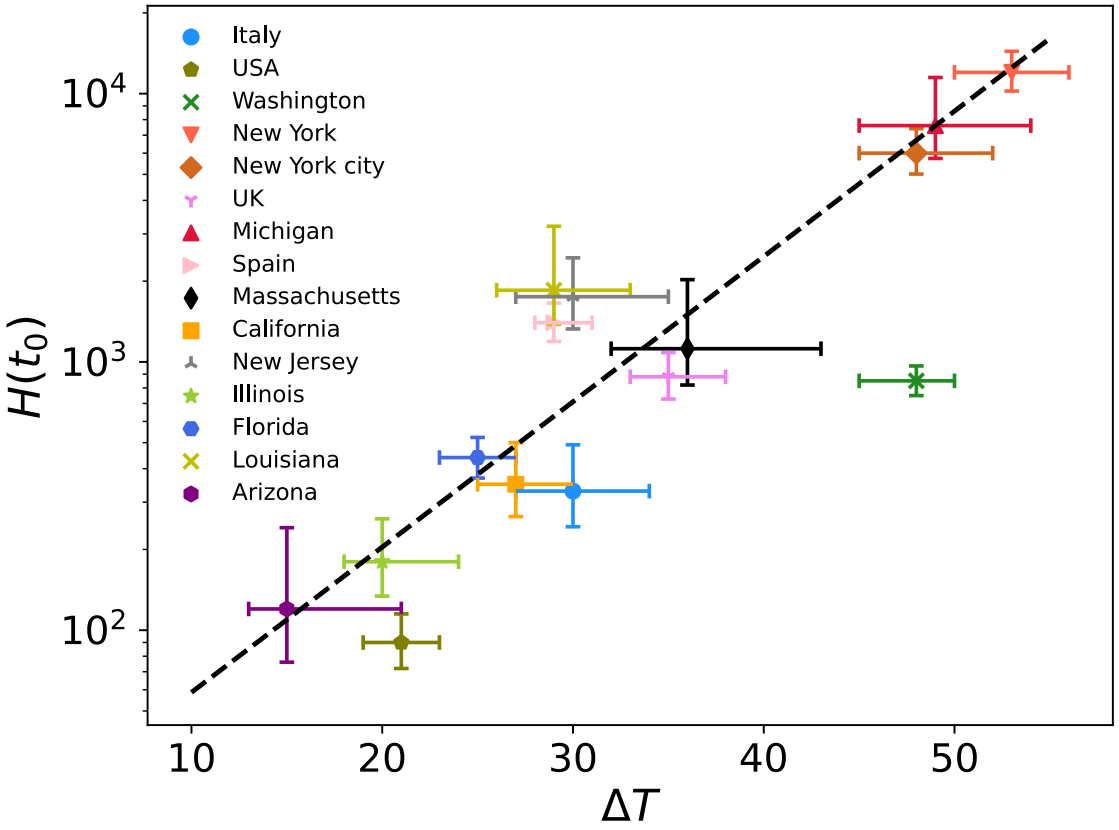

FIG. 4. Approximately exponential relation between the size of the hidden population at the time of first confirmed case reported and the time elapsed since ZERO. For the 15 systems studied, both the duration ΔT and the hidden population $H(t_0)$ vary widely. Among all the systems, when the first confirmed case was reported, the hidden population ranges from hundreds to thousands.

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 zero 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 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 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 US 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 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).

DISCUSSION

There were speculations that the novel coronavirus could have been in the US a few weeks earlier than January 20, the 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 US 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. Firstly, 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. Secondly, the key model parameters are optimally estimated using empirical data from an adequately long time period including the initial growth phase of the epidemic. Thirdly, 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 ten US States and New York city, the US 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/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.

METHODS

In the US, 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/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].

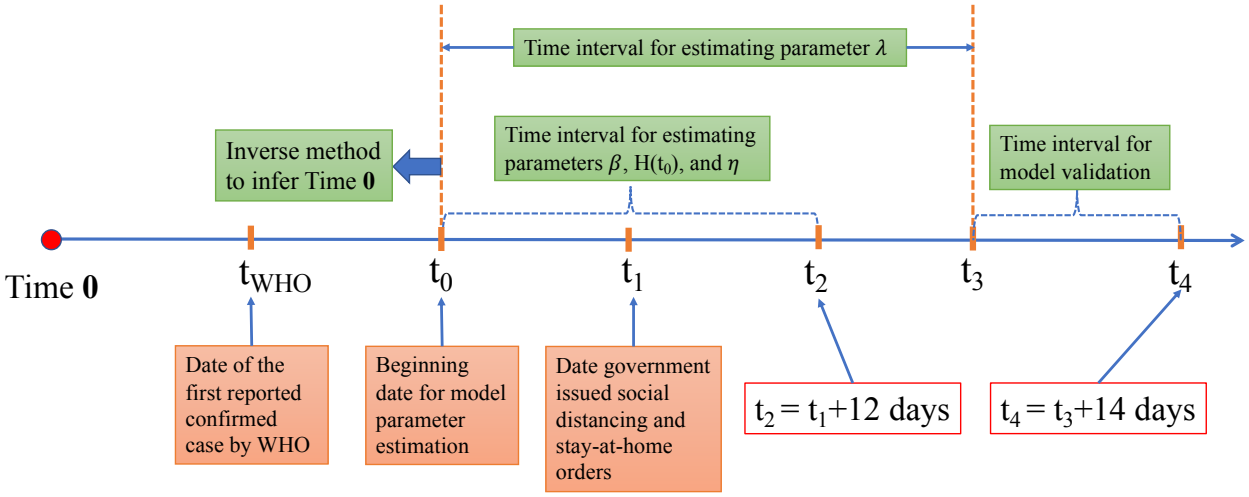

FIG. 5. Dates and timeline of inferring time **0** for a given State or country. The time t_{WHO} is the date of first emergence of case(s) reported by the World Health Organization (WHO). Time t_0 is the date when the number of confirmed cases began to increase, which is used as the initial time for 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 Supplementary Table 1 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: (β, $H(t_0)$, and η), 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.

To determine time **0** for a State in the US 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 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 as a Supplementary Note in Supplementary Information (SI).

Figure 5 explains our principle to infer time **0** in terms of a number of key dates underlying COVID-19 spreading. Some days after time **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: (β, $H(t_0)$, η) 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 (β, $H(t_0)$, η). 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 (β, $H(t_0)$, η) to yield an equal number of combinations (β, $H(t_0)$, η, λ). 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. 5. 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 $H(t_0)$: the starting date that gives the correct $H(t_0)$ value is taken to be time $\mathbf{0}$.

DATA AND CODE AVAILABILITY

All relevant data are available from the authors upon request. All relevant computer codes are available from the authors upon request.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 11975099, 11575041, 11675056 and 11835003), the Natural Science Foundation of Shanghai (Grant No. 18ZR1412200), and the Science and Technology Commission of Shanghai Municipality (Grant No. 14DZ2260800). YCL 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 Engineering and funded by the Office of Naval Research through Grant No. N00014-16-1-2828.

AUTHOR CONTRIBUTIONS

M.T. and Y.-C.L. designed research; Z.-M.Z. and Y.-S.L. performed research; Z.-M.Z. and Y.-S.L. contributed analytic tools; Y.-S.L., Z.-M.Z., M.T., Z.L., and Y.-C.L. analyzed data; Y.-C.L. and M.T. wrote the paper.

COMPETING INTERESTS

The authors declare no competing interests.

CORRESPONDENCE

To whom correspondence should be addressed. E-mail: tangminghan007@gmail.com; Ying-Cheng.Lai@asu.edu

---

[1] Gonzalez-Reiche, A. S. *et al.* Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* (2020).

[2] Gharavi, E., Nazemi, N. & Dadgostari, F. Early outbreak detection for proactive crisis management using Twitter data: COVID-19 a case study in the US. *arXiv:2005.00475* (2020).

[3] Long, Y.-S. *et al.* Quantitative assessment of the role of undocumented infection in the 2019 novel coronavirus (COVID-19) pandemic. *arXiv:2003.12028* (2020).

[4] Fauver, J. R. *et al.* Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. *Cell* **181**, 990 – 996.e5 (2020).

[5] Kraemer, M. U. G. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).

[6] Wu, J. T. *et al.* Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med.* 1–5 (2020).

[7] Li, Q. *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New Eng. J. Med.* (2020).

[8] Guan, W.-J. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *New Eng. J. Med.* (2020).

[9] Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. A novel coronavirus outbreak of global health concern. *Lancet* **395**, 470–473 (2020).

[10] Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet* **395**, 689–697 (2020).

[11] Zhou, T. *et al.* Preliminary prediction of the basic reproduction number of the Wuhan novel coronavirus 2019-nCoV. *J. Evidence-Based Med.* (2020).

[12] Du, Z. *et al.* The serial interval of COVID-19 from publicly reported confirmed cases. *medRxiv* (2020).

[13] Ferguson, N. M. *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.

[14] Flaxman, S. *et al.* Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries (2020). Report of Imperial College COVID-19 Response Team, March 30, 2020.

[15] Aleta, A. & Moreno, Y. Evaluation of the incidence of COVID-19 and of the efficacy of contention measures in Spain: a data-driven approach. *medRxiv* (2020).

[16] Rabajante, J. F. Insights from early mathematical models of 2019-nCoV acute respiratory disease (COVID-19) dynamics. *arXiv:2002.05296* (2020).

[17] Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* (2020).

[18] Klein, B. *et al.* Assessing changes in commuting and individual mobility in major metropolitan areas in the United States during the COVID-19 outbreak. *Report* (2020).

[19] Radulescu, A. & Cavanagh, K. Management strategies in a SEIR model of COVID 19 community spread. *arXiv:2003.11150* (2020).

[20] Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dise.* (2020).

[21] Pei, S. & Shaman, J. Initial simulation of SARS-CoV2 spread and intervention effects in the continental US. *medRxiv* (2020).

[22] Liu, P., Beeler, P. & Chakrabarty, R. K. COVID-19 progression timeline and effectiveness of response-to-spread interventions across the United States. *medRxiv* (2020).

[23] Stier, A. J., Berman, M. G. & Bettencourt, L. COVID-19 attack rate increases with city size. *arXiv:2003.10376* (2020).

[24] Lu, M. Dynamic modeling COVID-19 for comparing containment strategies in a pandemic scenario. *arXiv:2003.13997* (2020).

[25] Maier, B. F. & Brockmann, D. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science* **368**, 742–746 (2020).

[26] Tian, H. *et al.* An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **368**, 638–642 (2020).

[27] Hsiang, S. *et al.* The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* (2020).