scientific reports

Check for updates

OPEN

Identify hidden spreaders of pandemic over contact tracing networks

Shuhong Huang[1,7,8], Jiachen Sun[2,8], Ling Feng[3,4], Jiarong Xie[1], Dashun Wang[5] & Yanqing Hu[6]✉

The COVID-19 infection cases have surged globally, causing devastations to both the society and economy. A key factor contributing to the sustained spreading is the presence of a large number of asymptomatic or hidden spreaders, who mix among the susceptible population without being detected or quarantined. Due to the continuous emergence of new virus variants, even if vaccines have been widely used, the detection of asymptomatic infected persons is still important in the epidemic control. Based on the unique characteristics of COVID-19 spreading dynamics, here we propose a theoretical framework capturing the transition probabilities among different infectious states in a network, and extend it to an efficient algorithm to identify asymptotic individuals. We find that using pure physical spreading equations, the hidden spreaders of COVID-19 can be identified with remarkable accuracy, even with incomplete information of the contract-tracing networks. Furthermore, our framework can be useful for other epidemic diseases that also feature asymptomatic spreading.

As the COVID-19 pandemic continues to spread at rapid rates[1–3], and the development of effective pharmacological treatments is still uncertain according to WHO, non-pharmacological interventions like isolation of the infectious through quarantines[4,5] are the most effective and possibly the only means of containing the continued outbreaks, as it effectively reduces the person to person transmissions [6]. Yet, unlike other infectious diseases like SARS and Ebola, COVID-19 is unique in that a large portion of its infected population is mild or asymptotic[7]. Even some of the asymptotic infections do not exhibit any clinic symptoms until self-recovery[8,9]. Without being detected and subsequently quarantined, the asymptomatic population (i.e. hidden spreaders) sustains the ongoing spreading of the disease to the susceptible population unknowingly[10,11]. This poses a major challenge in the effective mitigation of the pandemic spreading. Furthermore, empirical studies have shown that such asymptomatic infections accounts for a large proportion of the population[12–18], as much as up to 80%[18]. Currently, estimation of the asymptomatic cases is done through exhaustive screening of close contacts of the known infected cases in the contact tracing networks[17]. This untargeted method requires large amount of resources and is time consuming, that in turn leads to ineffective or delayed interventions to quarantine the asymptomatic cases. On the other hand, the combination of a mobile app-based contact tracing network[19] and a statistical framework[20] shows the potential to accurately localize high-risk spreaders[21,22]. Hence, a targeted screening in the contact tracing network is pertinent, such that asymptomatic individuals can be estimated with high precision for intervention and spreading mitigation.

Here we incorporate the empirical characteristics of the COVID-19 spreading dynamics into a Markovian process, i.e. vectors that represent the different infection stages and their associated transition probabilities. By embedding the transition process into a contact tracing network that includes the known infected nodes (individuals), we develop a method that predicts the infectious states of the rest of the network with high precision. By combining such predictions with the network structure, we then derive the spreading power of every node

[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China. [2]Tencent, Shenzhen 518057, China. [3]Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138632, Singapore. [4]Department of Physics, National University of Singapore, Singapore 117551, Singapore. [5]Kellogg School of Management, Northwestern University, Evanston, IL, USA. [6]Department of Statistics and Data Science, College of Science, Southern University of Science and Technology, 518055 Shenzhen, China. [7]Present address: Institute of Neuroscience, Technical University of Munich, Munich 80802, Germany. [8]These authors have contributed equally: Shuhong Huang and Jiachen Sun. ✉email: yanqing.hu.sc@qq.com

taking into account of both its infectious state and its specific location in the network, such that screening of the asymptomatic can be prioritised accordingly. The effectiveness of our method is validated by empirical data from two COVID-19 transmission networks in Singapore. Moreover, in the simulated COVID-19 transmission experiment of contact-tracing network, we find that a screening scheme designed by the proposed computational framework outperforms several machine-learning baselines designed in this work and the random screening of infection neighbors. The latter was widely used in early COVID-19 outbreaks in China. Furthermore, even in the realistic situation of incomplete information on the contact tracing network, with missing links or sub networks consisting of only contacts of the infected cases, our method retains high accuracy. Thus our method is highly effective in asymptomatic case estimation and can be implemented to any contact-tracing networks either constructed manually[23,24] or through technological means[25] such as Bluetooth[26,27], GPS[28] and digital check-in check-out technologies (e.g. health QR codes[29] widely used in China).

Given the spreading of the COVID-19 occurring over the contact network, the challenge is to identify asymptomatic nodes with the information of infected symptomatic individuals (nodes) that have been identified from a certain time T. We approach this by estimating the probability of each node being in the infected state as illustrated in Fig. 1. Specifically, we first construct the transition dynamical equations among different infection stages and states based on the empirically observation of COVID-19 disease progression. The set of transition equations is then combined with the contact network topology and data on the observed infection history to deduce the state of each node in the network.

As observed in many clinical studies, the hidden spreaders of COVID-19 fall into two different categories. One is the presymptomatic infections who are asymptomatic and infected, but will later develop clinical symptoms (e.g. fever, cough, dyspnea, etc.); The other type corresponds to the asymptomatic patients who carry the virus but have never exhibited any symptoms until recovery. As a result, an individual can have a total of 5 different

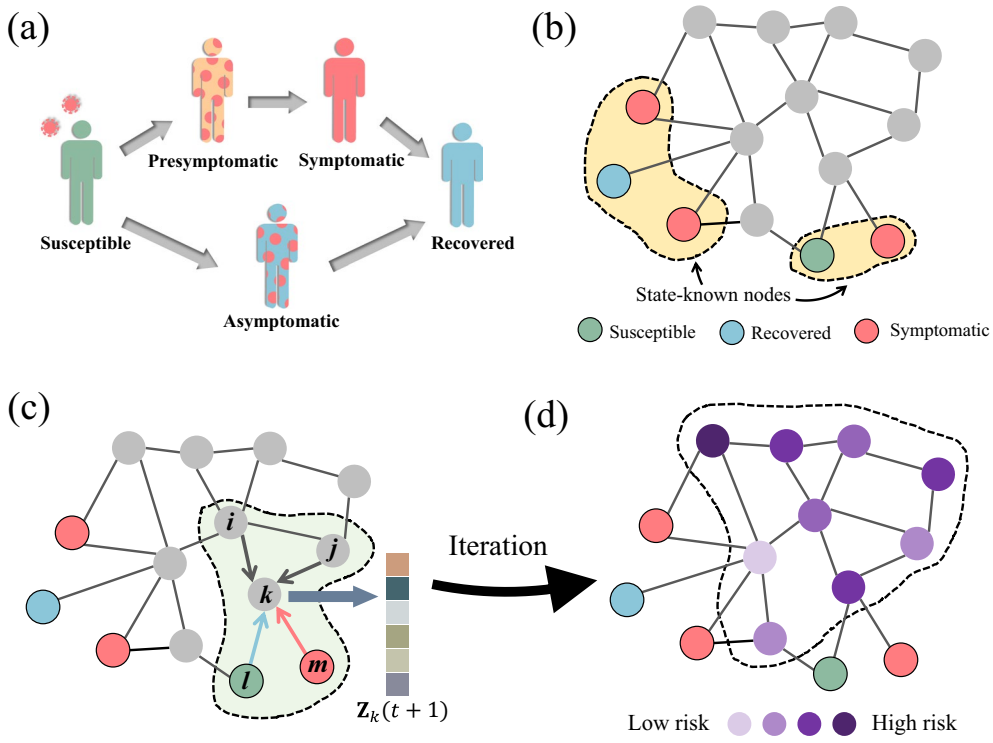

Figure 1. Identifying Asymptomatic and Presymptomatic COVID-19 Infections on Contact-tracing Networks. (**a**) The COVID-19 state transition of an individual. A susceptible will become either asymptomatic or presymptomatic after being infected. An asymptomatic patient will further turns to the symptomatic state after an incubation period, while an presymptomatic patient will never exhibit any symptoms until the recover. (**b**) Illustration of asymptotic/presymptomatic node identification problem. In a contact-tracing network, only a small fraction of the nodes' states are known (marked with color), while the hidden asymptomatic/presymptomatic individuals within the population (marked with grey) are potential spreaders. Our purpose is to find asymptomatic/presymptomatic infected individuals in the population using the contact-tracing network and the information of known confirmed cases. (**c**) Diagram of the proposed method. The state transition of an unknown node k is modeled as a Markov process, i.e., a vector $\mathbf{Z}_k$ where the elements represent the probabilities of different infection stages in (**a**). The specific value of the vector $\mathbf{Z}_k(t+1)$ at $t+1$ is determined by the infection status of known nodes at t and the structure of the contact network. (**d**) After iterations over the whole network, each unknown node will be assigned with an infection indicator according to the eventual values of its state vector $\mathbf{Z}_k$, which represents the risk of being infected.

states in the process of COVID-19 spreading (see Fig. 1a), namely, Susceptible (S), Presymptomatic (P), Asymptomatic (A), Symptomatic Infectious (I) and Recovered (R). Since the infectious duration in the states of P, I and A follows a specific probability distribution, here we further break down the P, I and A states into finer states representing the progression in each of the 3 states, i.e., the number of days passed since the beginning of the states. For better clarity, we denote t as the number of days of the COVID-19 evolution on the entire network and d as the number of days in a particular infected state for a particular individual. Since an individual i can be at any stage in the process, we can use $\mathbf{Z}_i(t)$ to represent the state probabilities at time t:

$$\begin{aligned}\mathbf{Z}_i(t) =&(S_i(t), P_i^1(t), ..., P_i^d(t), ..., I_i^1(t), ...,\\ &I_i^d(t), ..., A_i^1(t), ..., A_i^d(t), ..., R_i(t))\end{aligned} \tag{1}$$

where $S_i(t)$ and $R_i(t)$ is the probability that the individual i is susceptible and recovered at day t, respectively. $P_i^d(t)$, $I_i^d(t)$ and $A_i^d(t)$ are the probabilities that i is in the state of P, I and A for d days at the time of t. Since all of asymptomatic, presymptomatic and symptomatic states are infected states, their total probability corresponds to that of a node is infectious, and we use $C_i(t)$ to represent it:

$$C_i(t) = P_i(t) + I_i(t) + A_i(t) \tag{2}$$

where $P_i(t) = \sum_{d=1}^{\infty} P_i^d(t)$, $I_i(t) = \sum_{d=1}^{\infty} I_i^d(t)$, $A_i(t) = \sum_{d=1}^{\infty} A_i^d(t)$. Throughout this work we use $C_i(t)$ as a key indicator to infer whether an individual is infected .

From here, we can extract the probability transition dynamics among the 5 different states as follows. First, for a node who is in the susceptible state S at t, its next state at $t+1$ will be jointly determined by the state of its neighbors in the network at t. Specifically, the probability of a node i in S state remains in S on day $t+1$ (i.e., not infected by any of its infected neighbor on the next day) is:

$$S_i(t+1) = S_i(t) \cdot \prod_{j \in \partial i} (1 - \mathcal{F}(t, j, \beta)) \tag{3}$$

where ∂i represents the set of neighbors (contacts) of i in the network, $\mathcal{F}(t, j, \beta)$ represents the probability that i is infected by j. Since Eq. (3) origins of belief propagation, the infection possibility is strictly accurate only if the network has a tree structure[30]. However, Eq. (3) can still provide reasonable results on many networks with loops[31–33]. Threfore, the infection of node i can only happen if j is in the infected state on day t (probability $C_j(t)$), and happens to transmit it to i (probability β). Then we have:

$$\mathcal{F}(t, j, \beta) = C_j(t) \cdot \beta \tag{4}$$

Here β can be estimated from the empirically observed disease reproduction number R_0 for COVID-19 and the average number of neighbors in the contact tracing network $\langle k \rangle$. Specificaly, $\beta = \frac{R_0}{\lambda \langle k \rangle}$, where λ is the average time a susceptible person carries the virus, which can be expressed as $\lambda = p \cdot \mu_A + (1-p) \cdot \left(\exp\left(\mu_P + \frac{\sigma_P^2}{2}\right) + \mu_I \right)$, where p is the proportion of asymptomatic infected cases, $\mu_A, \exp\left(\mu_P + \frac{\sigma_P^2}{2}\right), \mu_I$ are the average time of the virus carried by infected individuals in A, P and I states[34,35] respectively.

Next, for an individual under S state at time t, the probability of becoming presymptomatic state P at $t+1$ is:

$$P_i^1(t+1) = S_i(t) \cdot (1-p) \cdot \left(1 - \prod_{j \in \partial i} \left(1 - \mathcal{F}(t, j, \beta)\right)\right) \tag{5}$$

Accordingly, we can calculate the probability that the state of i become A at $t+1$ as:

$$A_i^1(t+1) = S_i(t) \cdot p \cdot \left(1 - \prod_{j \in \partial i} (1 - \mathcal{F}(t, j, \beta)\right) \tag{6}$$

In the third case where a node i is in the infected state (i.e. E, I or A, $d \geq 1$) on day t, the transition probabilities that they will stay in the same state on day $t+1$ are:

$$\begin{aligned} P_i^{d+1}(t+1) &= P_i^d(t) \cdot \left(\frac{1 - F_P(d)}{1 - F_P(d-1)}\right)\\ I_i^{d+1}(t+1) &= I_i^d(t) \cdot \left(\frac{1 - F_I(d)}{1 - F_I(d-1)}\right)\\ A_i^{d+1}(t+1) &= A_i^d(t) \cdot \left(\frac{1 - F_A(d)}{1 - F_A(d-1)}\right)\end{aligned} \tag{7}$$

where $F_P(d) = \int_{-\infty}^{d} f_P(t)dt$, $F_I(d) = \int_{-\infty}^{d} f_I(t)dt$, $F_A(d) = \int_{-\infty}^{d} f_A(t)dt$ are the cumulative distribution functions of duration length d for P, I, A states, respectively. For mathematical convince, we simply set $F_P(0) = F_I(0) = F_A(0) = 0$. The fourth case is that individual in the presymptomatic state P turns into the symptomatic infectious state I at the next day, and can be described with the following transition probability:

$$I_i^1 = \sum_{d=1}^{\infty} P_i^d \cdot \frac{F_P(d) - F_P(d-1)}{1 - F_P(d-1)} \tag{8}$$

In the fifth case, an individual in the state I or the state A has a certain probability of being recovered i.e, turning into the R state on the next day. From the above equation, we obtain the probability that the individual i is in the state of R at the time $t+1$ is:

$$\begin{aligned} R_i(t+1) =& R_i(t) + \sum_{d=1}^{\infty} A_i^d \cdot \frac{F_A(d) - F_A(d-1)}{1 - F_A(d-1)} \\ &+ \sum_{d=1}^{\infty} I_i^d \cdot \frac{F_I(d) - F_I(d-1)}{1 - F_I(d-1)} \end{aligned} \tag{9}$$

To validate our mathematical framework, we test it on a real contact-tracing network in the *Infectious Stay Away* exhibition[36] (ISA network, see Sect. SI 1 for data detailed description) with 410 individuals and average degree $\langle k \rangle$ of 13 (more experiments on another social network are illustrated in Sect. SI 5). We simulate the spreading with the empirically observed parameters on COVID-19 spreading mechanisms[34] (see Methods for the simulation details). All parameters used in simulation are listed in Table 1. From repeated simulations, we then obtain the probability of every possible state of a node, and compare this baseline with the theoretical results from Eqs. (3–9). Here we set the dimension of Z to 77 according to the empirical temporal distributions of the infected states[34,35] (see Methods for detail). From Fig. 2a and b, we can see that our theoretical result on the temporal evolutions of the disease in the whole network is well validated by the simulations. These show that our transition probability framework is accurate in producing the real spreading dynamics.

Now we extend the proposed transition probability equations to identify nodes with high risk of being asymptomatic, assuming the infection history on symptomatic nodes is already known. The underlying principle is to update every node's state by incorporating the information of known infection into Eqs.(2–9) in the subsequent days, and then deduce the infection probability $C_i(T)$ for each node i in the network (see the details in the Methods). The nodes with higher $C_i(T)$ are identified as having high risk of being infected at day T. We test the effectiveness by applying it on two sets of real COVID-19 spreading data on the contact-tracing network in Singapore[23,24] (see Fig. 2c and d). The details of network is provided in Method and in Sect. SI 1. We find that the ranking our $C_i(T)$ values are highly correlated with the date of infection t of nodes (Fig. 2e and f), meaning nodes with higher infection probabilities indeed have higher risk of being infected in the real COVID-19 spreading data.

The Singapore empirical datasets have the constraint of merely including the symptomatic individuals' identities in the network. Therefore, to further evaluate our method, we simulate a realistic COVID-19 spreading process on the ISA network for T days to obtain the detailed infection history of every node in the network, such that the exact infection history on the asymptomatic nodes can be obtained. Assuming only the symptomatic nodes with state I are observed, i.e. infection histories of these nodes are known, we use our above method to identify those infected individuals among the rest of the nodes. Specifically, we select the nodes with the highest $C_i(T)$ values as the mostly likely infected nodes. In practice, the historical information of patients may be inaccurate based on personal statement. The infection time also needs to be inferenced from the dynamic structure of the contact tracking network. Inaccurate historical information may affect the perfomance. To our best knowledge, there is few prior works for estimating asymptomatic nodes in the network. Therefore, we also design several screening baselines based on the popular graph neural networks methods including Node2Vec[37], graph convolutional network [38] and graph attention networks[39] to further compare our results. Here the neural network models are utilized for an infected-uninfected binary classification task based on the network structure. A better perfomance can be achieved with more complex models. (Detailed methodologies for those methods in Sect. SI3).

The simulations results show that our transition probabilistic method (i.e. static screening) significantly outperforms the other methods in terms of the accuracy and recall on the local network where one can only observe the nearest neighbors of the known nodes in states I (see SI Fig. 1). Such advantage is still evident when we consider the alternative scenario that one can observe the full network structure[26,27] (see SI Fig. 3), and the intermediate scenario when only nearest and second nearest neighbors are known in the network (see SI Fig. 2). In a more realistic setting, the screening of the contact tracing network happens continuously in time. Here one can update the set of known infected nodes after every screening, and subsequently update the infected risk for the rest of the network from time to time. Therefore, we develop a dynamic screening method by updating the evaluation of $C_i(t)$ every time a new infected node is found through selective screening of the network (see the details in the Methods Section). This dynamic screening method outperforms (see Fig. 3a–d) other screening methods and even our previous static screening method (see Fig. 3d inset), implying that such dynamic screening method is highly effective in identifying infected nodes by screening less people.

Very often, the contact tracing network collected through either manual survey or digital tracking is at best incomplete, such that it is important to have a screening method that is still robust when there is missing information on the network structure. To test such robustness of our method, we randomly remove up to 80% of the edges in the ISA network, and test the accuracy of the method based on the remaining network (see the results on another network in SI Figs. 12–17). We find that the dynamic screening method on the various scenarios can still reliably identify the infected nodes in terms of accuracy and recall rate, as shown in Fig. 4 (see the robustness result on the static screening method in SI Fig. 16).

Lastly, we study the effectiveness of our method in containing the overall spread of COVID-19. In the widespread of COVID-19, limited resource on screening constrains the number of individuals the government can screen in a given day. Hence, targeted screening and mitigation can have significant impact on 'flattening the

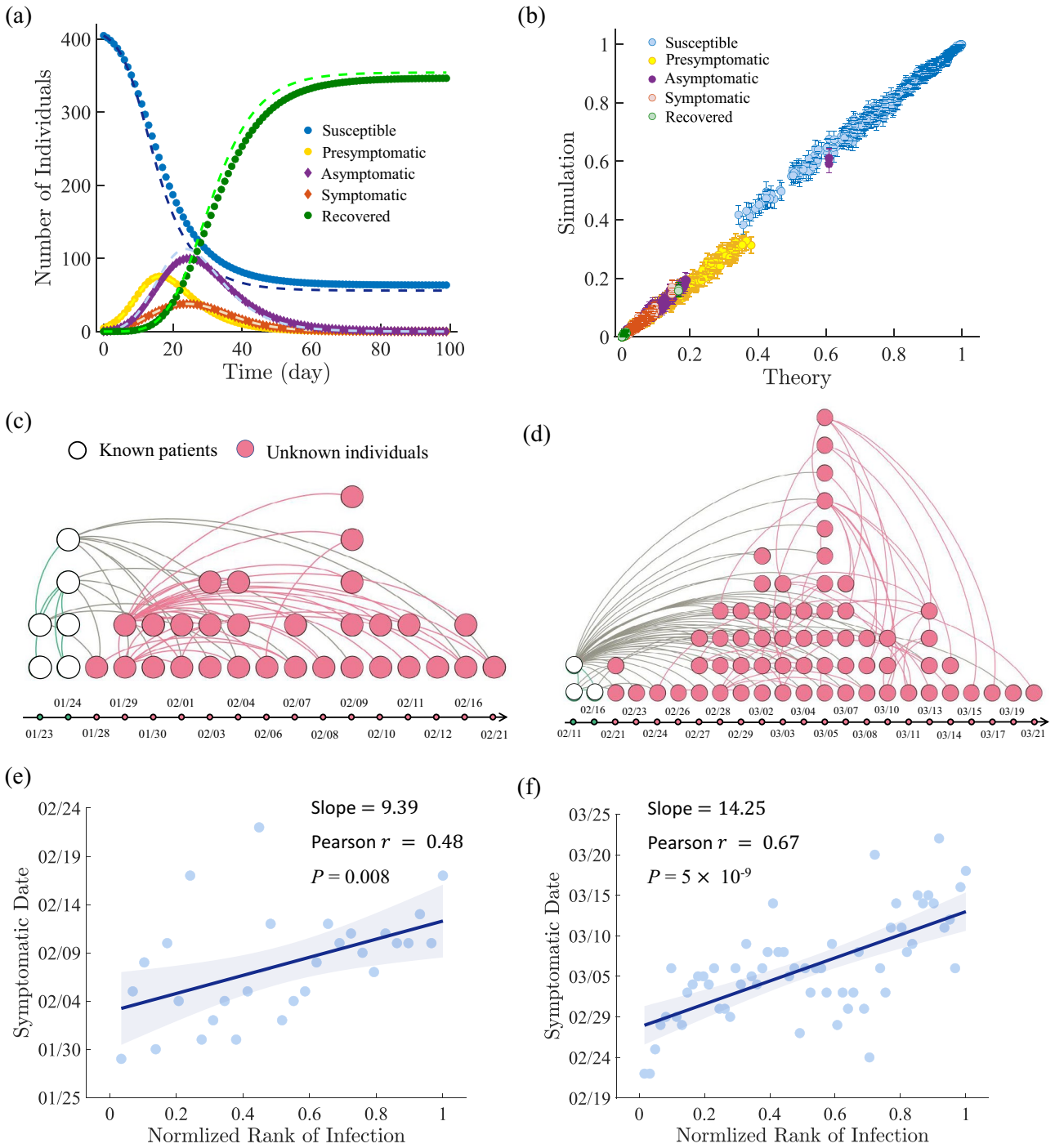

Figure 2. Empirical and Simulated Validation of the Proposed Model. (**a**) The number of people in each of the 5 states in the simulation process of COVID-19 spreading on the ISA network. At $T = 0$, we select three nodes with the maximum degree in the ISA network as the initial infected spreaders. Dash lines represent the theory values calculated by Eqs. (6–9). Dots represent the average value of 1000 simulations. (**b**) The theoretical probability vs. numerical frequency of each individual being in various states on $T = 10$ days. Each dot corresponds to a certain state of a node in the ISA network while the errorbar is the 95% confidence interval obtained by the bootstrapping method[41]. (**c**) The topological structure of a real COVID-19 spreading network in Singapore (Singapore A), where dots are patients and curves are contacts between patients (see Sect. SI 1 for the description of the network). The points on the timeline indicate the date of the patient's presence. (**d**) The topological structure of another network of Singapore (Singapore B). (**e**) Relationship between the individual symptomatic time and the estimated infection probability in Singapore A. The network has a total of $T = 30$, from January 23, 2020 to February 21, 2020. Here we utilize the information of infections from the first two different time points as known set to infer the rest nodes' states in the network by Eqs. (1–9). Using the obtained state vector of each unknown node, we rank them according to the infection probability and compare with it's real symptomatic time. Since all patients are symptomatic in the dataset, the rank is based on $C_i(t) - A_i(t)$. The line denotes the linear fitted result and the shaded area denotes the 95% confidence interval. (**f**) Similar to (**e**), the value of the rank of infection probability versus symptomatic time in Singapore B. The network has a total of $T = 40$, from February 11, 2020 to March 21, 2020. We use infections who got infected the first two different time points for training.

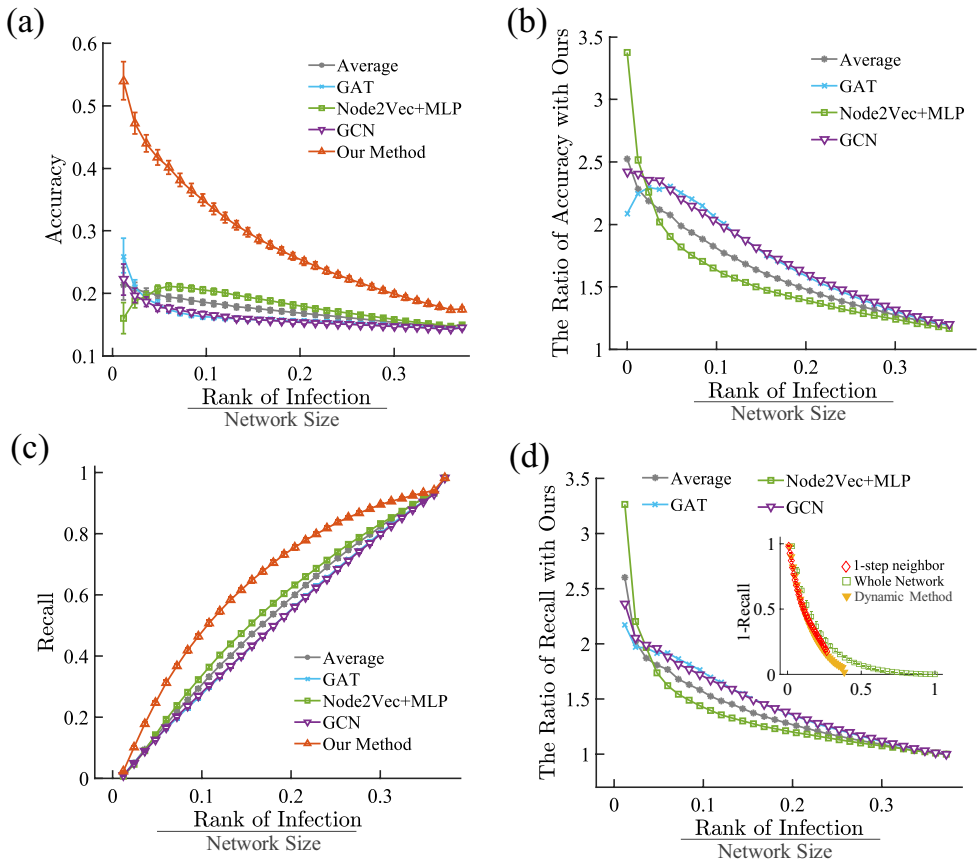

Figure 3. Screening Performance Assessment. We measure the performance of the dynamic screening method on the ISA network compared with other machine-learning baselines (see details in Sect. SI 3). (**a**) The accuracy vs. rank of infection (divided by the network size). The accuracy is defined as the proportion of non-Susceptible individuals in the ranking list. Since all nodes we screen at T do not have symptoms, here we use the value of $C_i(t) - I_i(t)$ to rank these nodes. (**b**) Similar to (**a**), the relative accuracy of the machine-learning-based algorithms, which is the ratio between the accuracy of our proposed algorithm and other algorithms. (**c**) The relationship between the rank of infection and the recall rate (i.e., the proportion of successfully identified non-Susceptible individuals to those in the whole network. (**d**) Similar to (**c**), the relative recall rate of the machine-learning-based algorithms. (inset) Recall rate of the static algorithms on the 1-step neighbor subnetwork and on the whole network (see SI Figs. 12–15), compared with the dynamic algorithm.

curve' of daily infected cases. To study such effect, we again simulate the COVID-19 spreading on the ISA network[36], and start screening/testing from day 10 using our method (i.e. 'neighbor containment'). Each day 2% (4% in SI Figs. 18, 19) of the whole network are tested for the disease, and the positive ones are immediately quarantined, corresponding a transition to the state R (see details of the containment strategy in Sect. SI 4). As shown in Fig. 5a–h, our method is highly effective in suppressing the daily infection cases and total infection cases, outperforming both the baseline strategy of only quarantining the infected ones (labeled as 'infection containment') and the strategy randomly screening 2%N among the neighbors of the known infections (labelled as 'neighbor containment'), where N is the network size. In addition, we find that even with up to 80% missing links, our method is still robust enough to effectively suppress the spreading, close to that of knowing the full network structure. It shows that our method is expected to be highly effective in containing COVID-19 spread in practice.

In this paper, based on the transmission rule of COVID-19 and the underlying physical spreading equations, we for the first time studied the estimation of asymptomatic infections in the contact-tracing network, which is a current major concern in the prevention and containment of COVID-19 worldwide. We provided a complete computational framework of inferring latent infection on contact network. Based on this, we proposed a feasible method for optimal detection of latent infection in combination with nodal transmission ability in the network. We show that the COVID-19 transmission can be broken in a timely and efficient manner by the proposed method, which outperforms the direct contact screening, a typical method widely used in China. In addition, our simulation on a real contact network demonstrated that, this method is robust even with incomplete network information, demonstrating its effectiveness in practical scenarios.

With the theoretical model presenting in this paper, it is the first time to focus on the inference task on a network from a spreading perspective. However, there are two main realistic concerns of the proposed framework.

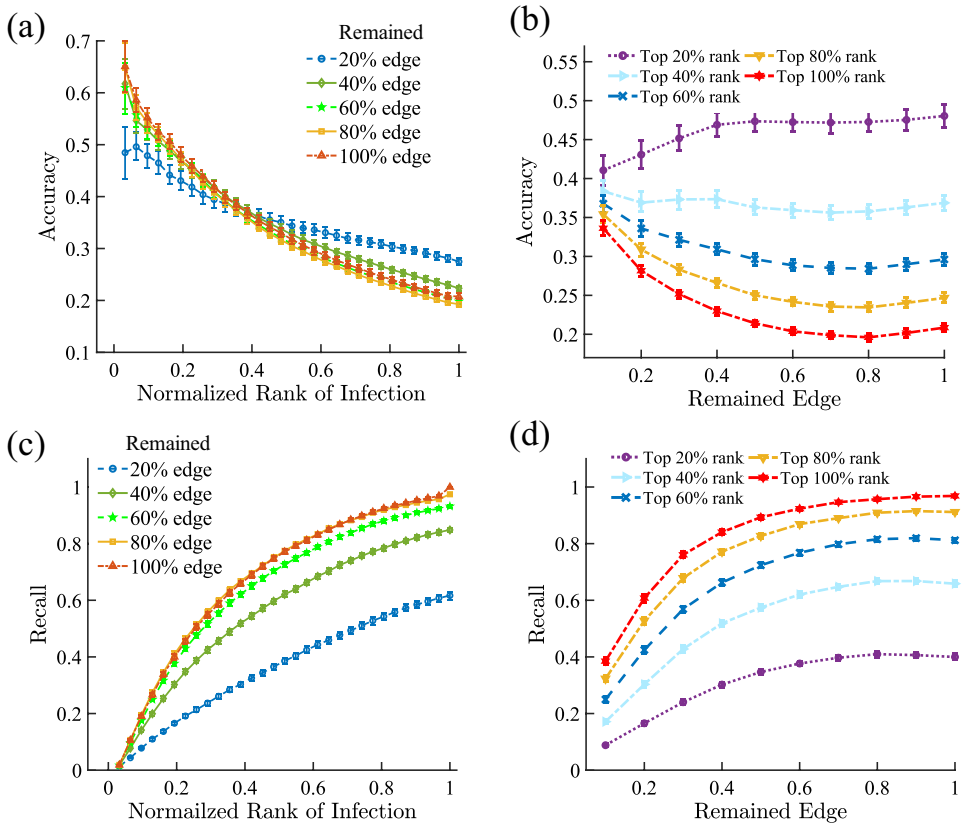

Figure 4. Performance of the Dynamic Screening Method with Incomplete Network Information. We randomly remove a fraction of links in the ISA network (see the result of another social network in Sect. SI 5) and then employ the proposed screening schemes on the remaining network. (**a**) The relationship between the accuracy and the ranking value of the infection probability with different proportions of the removed edges. Here we normalize the ranking value to range [0,1] by dividing it with the total number of individuals who have been screened. (**b**) Accuracy of the dynamic screening method versus the proportion of the removed edges by measuring the infection rank with different proportions. For example, top 20% rank means the 20% nodes with the highest infection possibility which is equal to 0.2 of normalized rank of infection in (**a**). (**c**) The dependencies of recall rate of the dynamic screening method on the infection rank. (**d**) Recall versus the remaining edges.

On the one hand, The virus transmission constant β may be varied for different virus mutant. As shown in the supplementary information, the model is robust among different transmission constant β. As the epidemic progress, updating the transmission constant dynamically could address this issue. On the other hand, the transmission ability has individual variances. For this complex situation, the proposed model can be improved to satisfy the practice. For example, the nodes on the network can be grouped based on their personal information such as age, past medical history, etc. For different individual groups, different transmission dynamics can be established.

Therefore, we believe that the theory and the corresponding methods in identifying COVID-19 hidden spreaders are of great practical significance. In principle, it provides policymakers and front-line workers in COVID-19 with important and effective guidance and tools that could be deployed swiftly to fight COVID-19, and save billions of people around the world who are still suffering as the epidemic continues to spread throughout the world.

Methods

Singapore COVID-19 datasets. The data was collected by the Singapore government[23,24], and contains comprehensive records on the dates of showing symptoms and confirming the disease, as well as their contact networks. We pick the infected nodes from the first two different time points, and set T to January 26, 2020 and February 19, 2020 for Singapore A and Singapore B, respectively. Then based on the known infection history of the nodes, our transition probability method estimates the infection probabilities $C_i(T)$ of every other node in the networks.

The dimension of the state vector. The dimension of the state vector $\mathbf{Z}_i(t)$. The dimension of the state vector $\mathbf{Z}_i(t)$ corresponds to the total number of sub states possible during the various disease progression paths, i.e., the number of days that an individual can be in each of the 3 different infected states. From the empirical

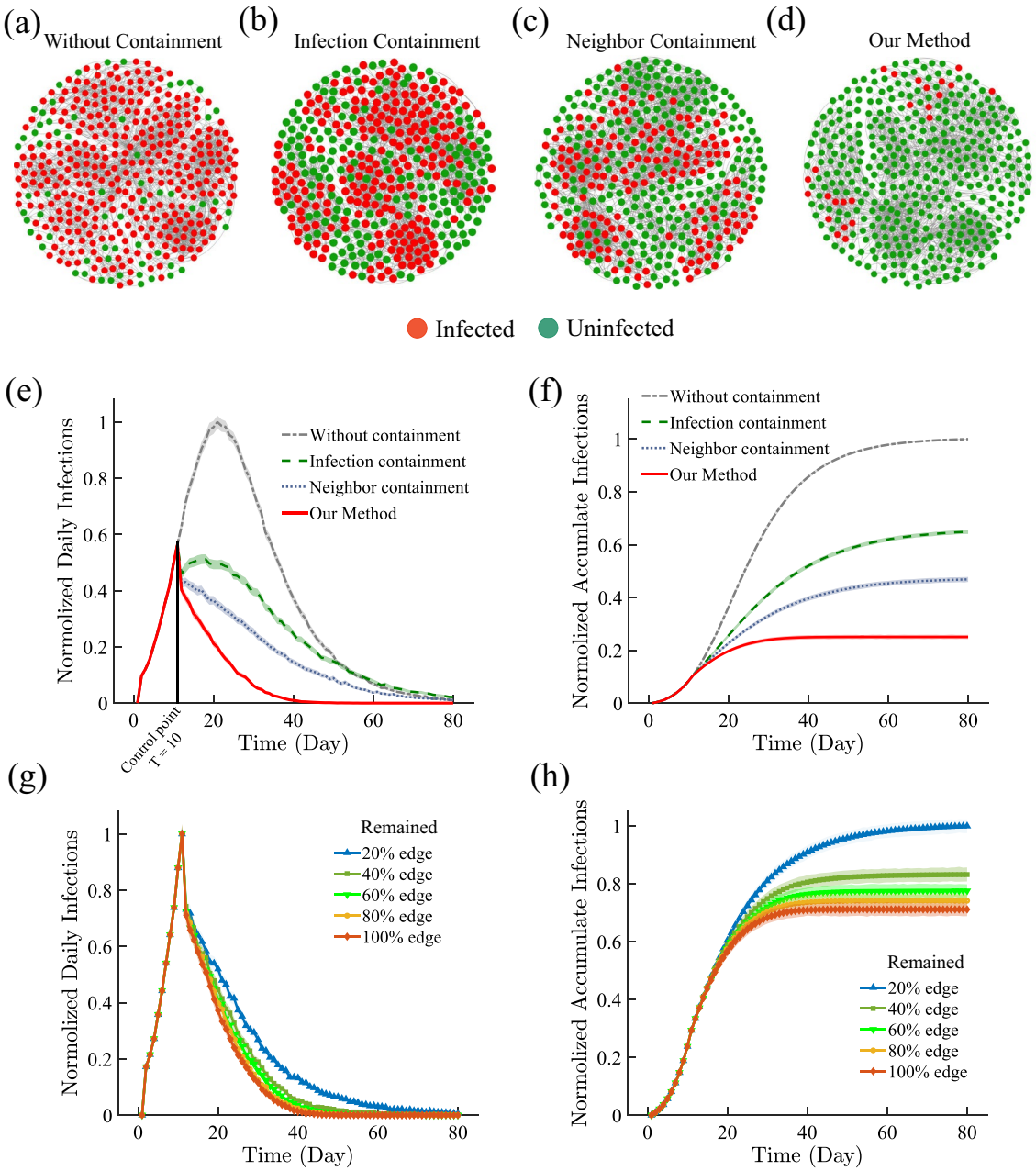

Figure 5. Containment Effectiveness of the Proposed Scheme on Simulated COVID-19 Spreading. From $T = 10$ of COVID-19 spreading simulation on the ISA network, we conduct different approaches separately to contain the pandemic, including the proposed method (i.e. Dynamic Containment), the Infection Containment and the Neighbor Containment (see Sect. SI 4 for details). For the Neighbor Containment and the proposed method, we select a total of 2%N of individuals to screen and then quarantine those tested positive at each time step. (**a**) Visualization of infected population without conducting any contain scheme and (**b**–**d**) the 3 different control schemes). (**e**) The number of daily new infections in different control schemes. The value of each curve is divided by the highest point at that time to normalize to [0,1]. (**f**) The cumulative number of infections corresponding to (**e**). (**g**–**h**) The performance of our control scheme under an incomplete network where a fraction of links are randomly removed in the detection process.

temporal distributions of the infected states[34,35] (Table 1), we use 3 standard deviations[40] as cut off on the max number of days in states P, I, A, which are 20, 20, 35 days, yielding a dimension of 77 for Z. (S and R states have no sub states).

COVID-19 spreading simulation. COVID-19 spreading simulation. At the starting time $T = 0$, we select the 3 nodes with the largest degree in the network as initial infected nodes, whose infected states are determined as either Asymptomatic or Presymptomatic according to the parameter p of Table 1. Then we apply the empiri-

Parameter	Meaning	Value	Origin
R_0	Basic reproduction number	3.50	Average from 10 researches[42–51]
p	Fraction of asymptomatic infections	15%	Minimal value from 5 researches[13,14,14,16,17]
$f_P(d)$	Distribution of during length of presymptomatic state	Logarithmic normal distribution with $\mu_P = 1.43$ and $\sigma_P = 0.66$	Fitted value from clinical data of[34]
$f_I(d)$	Distribution of during length of symptomatic state	Normal distribution with $\mu_I = 8.8$ and $\sigma_I = 3.88$	Fitted value from clinical data of[34]
$f_A(d)$	Distribution of during length of asymptomatic state	Normal distribution with $\mu_A = 20.0$ and $\sigma_A = 5.0$	μ_A is estimated from clinical data of[35]

Table 1. COVID-19's clinical parameters and infectious characteristics used in this work.

cally observed parameters on COVID-19 spreading mechanisms[34] including reproductive number $R_0 = 3.50$ and asymptotic infection ratio $p = 15\%$ on our equations to simulate the spreading. The set of values are listed in Table. 1(see Sect. SI 1 for the detail description of parameters and Sect. SI 3 for the discussion of the parameter sensitivity). Each simulation corresponds to one realization of the actual spreading based on the realistic dynamics, and the actual states of each node at every time step can be captured. More details of the simulation of COVID-19 are provided in Sect. SI 2.

Identifying infection probability. Identifying infection probability $C_i(T)$. The goal is to identify nodes with high risk of being asymptomatic with infection history on known symptomatic nodes, and we extend our transition probability equations to study this problem. At a certain time T, given the set I of infected individuals, the first day of infection s_j and the day of recovery r_j for each individual $j \in I$, we aim to develop a method from Eqs. (2–9) to deduce the infection probability $C_i(T)$ for each node i in the network. Note that the day of recovery can also be the day of death or quarantine. The initial condition at $t = 0$ is that every node in the network is in susceptible state, i.e. $\mathbf{Z}_i(0) = \{1, 0, \cdots, 0\}$. The day of first infection in the network is set to 1, i.e. $\min_{j\in I} s_j = 1$, and we update every node's state in the subsequent days depending on whether their infection history is known at time T. For the known nodes $j \in I$, we artificially assign their infection states according to the known information, meaning that j is assigned state S when $t < s_j$, state R when $t > r_j$, and infectious state when $s_j < t < r_j$. For the other nodes, we evaluate their state vector $Z_i(t)$ at every time step t according to the transition probabilities in Eqs. (2–9), until the final day T, such that their probabilities $C_i(T)$ of being infected can be evaluated from $Z_i(T)$.

Dynamic screening method. Dynamic screening method Every time we screen only node k that is of highest risk according to the algorithm; if node k is COVID-19 positive, it is added to the known infected nodes set I, and its neighbors are added to the unknown set, and we repeat the transition probability calculations according to Eqs. (2–9) from time $0 < t \leq T$; if k is negative, its probability state vector is set to be $\mathbf{Z}_k(t) = \{1, 0, \cdots, 0\}$ in the calculation of Eqs. (2–9). Next, the revised estimations of infection probabilities for each unknown node from Eqs. (2–9) tells us which node is the most risky and to be tested.

Data availability

All of network structures used in this paper and the code to simulate disease spreading can be found in https://github.com/Timbrer/HiddenSpreader.

Received: 11 October 2022; Accepted: 29 March 2023
Published online: 19 July 2023

References

1. Hui, D. S. *et al.* The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health-the latest 2019 novel coronavirus outbreak in Wuhan. *China Int. J. Infect. Dis.* **91**, 264 (2020).
2. Megan, S. How the pandemic might play out in 2021 and beyond. *Nature*(2020).
3. World Health Organization *et al.* WHO Director-general's opening remarks at the media briefing on COVID-19-11 March 2020"Who director-general's opening remarks at the media briefing on Covid-19-11 march 2020", (2020).
4. Hellewell, J. *et al.* Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Glob. Health* **8**(4), e488–e496 (2020).
5. Maier, B. F. & Brockmann, D. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science* **368**, 742 (2020).
6. Chan, J.F.-W. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. *Lancet* **395**, 514 (2020).
7. Gao, Z., Xu, Y., Sun, C., Wang, X., Guo, Y., Qiu, S. & Ma, K. A systematic review of asymptomatic infections with COVID-19 immunology and infection, J. Microbiol. (2020)
8. Kimball, A. *et al.* Asymptomatic and presymptomatic SARS-CoV 2 infections in residents of a long-term care skilled nursing facility-King County Washington. *Morb. Mort. Wkly. Rep.* **69**, 377 (2020).
9. Long, Q.-X. *et al.* Clinical and immunological assessment of SARS-CoV-2 asymptomatic infections. *Nat. Med.* **26**(8), 1200–1204 (2020).
10. Liu, Y.-C., Liao, C.-H., Chang, C.-F., Chou, C.-C. & Lin, Y.-R. A locally transmitted case of SARS-CoV-2 infection in TaiwanNew England. *J. Med.* **382**, 1070 (2020).
11. Rothe, C. *et al.* Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *N. Eng. J. Med.* **382**, 970 (2020).
12. Quilty, B. J. *et al.* Effectiveness of airport screening at detecting travellers infected with novel coronavirus (2019-nCoV). *Eurosurveillance* **25**, 2000080 (2020).

13. Byambasuren, O., Cardona, M., Bell, K., Clark, J., McLaws, M.-L. & Glasziou, P. Estimating the extent of true asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis, Available at SSRN 3586675 (2020).
14. Nishiura, H. *et al.* Estimation of the asymptomatic ratio of novel coronavirus infections (Covid-19). *Int. J. Infect. Dis.* **94**, 154 (2020).
15. Day, M. Covid-19 identifying and isolating asymptomatic people helped eliminate virus in Italian village. *Br. Med. J.*https://doi.org/10.1136/bmj.m1165 *(2020)*.
16. Yu, Y., Liu, Y.-R., Luo, F.-M., Tu, W.-W., Zhan, D.-C., Yu, G. & Zhou, Z.-H. medRxiv COVID-19 Asymptomatic Infection Estimation (2020)
17. Mizumoto, K., Kagaya, K., Zarebski, A. & Chowell, G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship Yokohama, Japan, 2020. *Eurosurveillance* **25**, 2000180 (2020).
18. Heneghan, C., Brassey, J., Jefferson, T. Centre for evidence-based medicine COVID-19: What proportion are asymptomatic? (2020).
19. Schneider, T. *et al.* Epidemic management and control through risk-dependent individual contact interventions. *PLOS Comput Biol.* **18**, e1010171 (2022).
20. Altarelli, F., Braunstein, A., Dall'Asta, L., Lage-Castellanos, A. & Zecchina, R. Bayesian inference of epidemics on networks via belief propagation. *Phys. Rev. Lett.* **112**, 118701 (2014).
21. Pei, S., Liljeros, F. & Shaman, J. Identifying asymptomatic spreaders of antimicrobial-resistant pathogens in hospital settings. *Proc. Natl. Acad. Sci.* **118**, e2111190118 (2021).
22. Baker, A. *et al.* Epidemic mitigation by statistical inference from contact tracing data. *Proc. Natl. Acad. Sci.* **118**, e2106548118 (2021).
23. LINKS ESTABLISHED BETWEEN CHURCH CLUSTERS AND WUHAN TRAVELLERS"Links established between church clusters and Wuhan travellers", (2020a),https://www.moh.gov.sg/news-highlights/details/links-established-between-church-clusters-and-wuhan-travellers,Lastaccessed4Apr2020.
24. "12 more cases discharged, 52 new cases of Covid-19 infection confirmed", (2020b),https://www.moh.gov.sg/news-highlights/details/12-more-cases-discharged-52-new-cases-of-covid-19-infection-confirmed,Lastaccessed4Apr2020.
25. Kondylakis, H. *et al.* COVID-19 mobile apps: A systematic review of the literature. *J. Med. Internet Res.* **22**(12), e23170 (2020).
26. Drew, D. A. *et al.* Rapid implementation of mobile technology for real-time epidemiology of COVID-19. *Science* **368**(6497), 1362–1367 (2020).
27. Ferretti, L. *et al.* Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, eabb6936 (2020).
28. Chang, S. *et al.* Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**(7840), 82–87 (2020).
29. Mozur, P., Zhong, R., Krolik, A. In coronavirus fight, China gives citizens a color code, with red flags, New York Times **1** (2020).
30. Mezard, M. & Montanari, A. *Information, physics, and computation* (Oxford University Press, Oxford, 2009).
31. Cantwell, G. T. & Newman, M. E. Message passing on networks with loops. *Proc. Natl. Acad. Sci.* **116**, 23398 (2019).
32. Frey, B.J., MacKay, D. A revolution: Belief propagation in graphs with cycles, Adv. Neural Inf. Process. Syst. **10** (1997).
33. Murphy, K., Weiss, Y., Jordan, M.I. Loopy belief propagation for approximate inference: An empirical study arXiv preprint arXiv:1301.6725 (2013).
34. Li, Q. *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**(1199), 1207 (2020).
35. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet* **395**(10229), 1054–1062 (2020).
36. Isella, L. *et al.* What's in a crowd? Analysis of face-to-face behavioral networks. *J. Theor. Biol.* **271**, 166 (2011).
37. Grover, A., Leskovec, J. Node2vec: Scalable feature learning for networks, In *booktitle Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016) pp. 855–864.
38. Kipf, T.N., Welling, M. Semi-supervised classification with graph convolutional networks, In: *booktitle 5th International Conference on Learning Representations (ICLR)* (2016).
39. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. & Bengio, Y. Graph attention networks, In *booktitle 6th International Conference on Learning Representations (ICLR)* (2017).
40. Wheeler, D.J. & Chambers, D.S. Understanding statistical process control, USPC (1992).
41. DiCiccio, T. J. & Efron, B. Bootstrap confidence intervals. *Stat. Sci.* **11**, 189–212 (1996).
42. Tang, B. *et al.* Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. *J. Clin. Med.* **9**, 462 (2020).
43. Riou, J. & Althaus, C. L. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus(2019-nCoV), December 2019 to January 2020. *Eurosurveillance* **25**, 2000058 (2020).
44. Zhao, S. *et al.* Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in china, from 2019 to 2020: A data-driven analysis in the early phase of theoutbreak. *Int. J. Infect. Dis.* **92**, 214 (2020).
45. Liu, T., Hu, J., Kang, M., Lin, L., Zhong, H., Xiao, J., He, G., Song, T., Huang, Q. & Rong, Z. *et al.*Transmission dynamics of 2019 novel coronavirus (2019-nCoV) BioRxiv (2020b).
46. Imai, N., Dorigatti, I., Cori, A., Donnelly, C., Riley, C. & Ferguson, N. "Report 2: Estimating the potential total number of novel coronavirus cases in Wuhan city, China. 22 january 2020-imperial college London. who collaborating centre for infectious disease modelling. mrc centre for global infectious disease analysis, j-idea, imperial college london, uk", (2020).
47. Cao, Z., Zhang, Q., Lu, X., Pfeiffer, D., Jia, Z., Song, H. & Zeng, D.D.Estimating the effective reproduction number of the 2019-nCoV in China, MedRxiv (2020).
48. Shen, M., Peng, Z., Xiao, Y. & Zhang, L. Modelling the epidemic trend of the 2019 novel coronavirus outbreak in China, BioRxiv (2020).
49. 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 (2020).
50. Read, J.M., Bridgen, J.R., Cummings, D.A., Ho, A. & Jewell, C.P. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions MedRxiv (2020).
51. Majumder, M. & Mandl, K. D. Early transmissibility assessment of a novel coronavirus in Wuhan, China, China (2020).

Acknowledgments

This work is partially supported by National Natural Science Foundation of China under Grants No. 12275118, Natural Science Foundation of Guangdong for Distinguished Youth Scholar, Guangdong Provincial Department of Science and Technology (grant no. 2020B1515020052), Guangdong High-Level Personnel of Special Support Program, Young TopNotch Talents in Technological Innovation (grant no. 2019TQ05X138) and NUS AcRF Grant A-0004550-00-00.

Author contributions

Y.H. conceived the project. S.H., Y.H., J.S., L.F., J.X., D.W. designed the experiments. S.H., performed the experiments and numerical modelling. Y.H., S.H., J.S., L.F., J.X., D.W., wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-32542-3.

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

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

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